HIV genome-wide diversity, interaction and coevolution

Guangdi Li

Promoter: Prof. Anne-Mieke Vandamme
Co-promoters: Prof. Jan Ramon, Dr. Kristof Theys

Doctoral thesis in Biomedical Sciences
Leuven, 2014

KU Leuven
Biomedical Sciences Group
Faculty of Medicine
Rega Institute for Medical Research
Department of Microbiology and Immunology
Laboratory of Clinical and Epidemiological Virology

HIV genome-wide diversity, interaction and coevolution

Guangdi Li

Promoter: Prof. Dr. Anne-Mieke Vandamme
Co-promoter: Prof. Dr. Jan Ramon
Dr. Kristof Theys
Chairman: Prof. Dr. Marc Van Ranst
Jury members: Prof. Dr. Marc De Maeyer
Prof. Dr. Hendrik Blockeel
Prof. Dr. Ann Nowé
Prof. Dr. Peter Cherepanov

Doctoral thesis in Biomedical Sciences

Leuven, 2014

Acknowledgement

A book is written for reading, an acknowledgement is craved for a lifetime memory. I would like to express my sincere gratitude to:

- My Phd promoter: Prof. Anne-Mieke Vandamme. This thesis could never have been completed without her consistent support during the last five years. Because of her encouragement and supervision, I can see the path to be a respectful and knowledgeable scientist.
- My Phd co-promoters: Prof. Jan Ramon and Dr. Kristof Theys. During my Phd study, Prof. Jan Ramon always provided valuable advice to improve my papers. I have also received extensive help from Dr. Kristof Theys.

I would like to acknowledge my collaborators who contributed to my research projects:

- Prof. Concha Bielza and Prof. Pedro Larrañaga, who have cultivated my research skills and have developed my mind as an independent thinker.
- Dr. Jens Verheyen generously shared his data and knowledge on HIV Gag.
- Dr. Arnout Voet inspired me with his ideas on drug design and has been an advisor on structural biology.

I would like to express my thanks to the jury members of my thesis

- Prof. Marc De Maeyer has been my mentor in structural biology.
- Prof. Hendrik Blockeel has inspired me on the development of data mining methods.
- Prof. Ann Nowé has provided advices on the improvement of my Phd thesis.
- Prof. Peter Cherepanov has been a hero and model for me to perform the outstanding HIV research.

I would like to express my appreciation for friendship and working experience to my colleagues in the whole laboratory:

- Prof. Philippe Lemey impressed me as a knowledgeable scientist, a kind friend and a thoughtful mentor.
- Prof. Marc Van Ranst is always a respectable leader and chief in our department.
- Prof. Ricardo Camacho has encouraged me his knowledge as an experienced doctor.
- Prof. Kristel Van Laethem gave me invaluable advice on HIV clinical care.

- Prof. Johan Van Weyenbergh shared with me his knowledge on immunology.
- Prof. Christophe Pannecouque always surprised me with his versatile talents.
- Prof. Dirk Daelemans always moved me with his passion for exciting research.
- Abbas Jariani shared with me many working nights in the lab.
- Andrea-Clemencia Pineda-Peña consistently encouraged me when facing difficulties.
- Annelies Wouters has been a kind friend cheering on every progress I made in lab.
- Bram Vrancken impressed me with his experience on deep sequencing.
- Britta Moens has enlightened many enjoyable moments in the lab.
- Carolina Alvarez showed me how to smile during the hardship.
- Eline Boons impressed me with her knowledge on antibody development.
- Elisabeth Heylen has always been a good friend.
- Filip Bielejec showed me his creativity on studying the spread of epidemic disease.
- Fossie Ferreira supported me a lot on data management and written English.
- Gertjan Beheydt helped me on the software development.
- Guy Baele advised and enlightened me on publication strategies.
- Jasper Edgar Neggers has always been a friend and has helped to improve my papers.
- Jelle Matthijnssens has been a good friend and knowledgeable scientist.
- Joao Sousa showed me his amazing expertise on the HIV origin and his skill on chess.
- Jurgen Vercauteren advised me on statistical tests.
- Kiyoshi Fukutani shared with me experiences in wet laboratory.
- Kris Covens shared with me the passion of doing research.
- Liana Eleni Kafetzopoulou helped me on the written English.
- Lies Laenen has always been working hard in lab.
- Lize Cuypers shared with me her insightful knowledge of hepatitis C virus.
- Lore Vinken shared with me her profound knowledge on viral sequencing.
- Mahmoudreza Pourkarim mentored me with his great knowledge on hepatitis B virus.
- Mark Zeller shared with me many advices on my Phd thesis.
- Mónica Eusébio impressed me with her skillful project managements.
- Nádia Conceição Neto deserves my considerable thanks for her help on my papers.
- Nídia Sequeira Trovão has been a good friend and expert on phylogenetic analysis.
- Nuno Rodrigues Faria helped me a lot on the phylogenetic analysis.

- Piet Maes has been a good friend and provided many advices.
- Pieter Libin impressed me with his experience on data management.
- Raphael Sangeda encouraged me to work on several research topics.
- Ria Swinnen and Sietse Huysmans helped me with official documents.
- Ricardo Khouri shared with me many of his brilliant ideas.
- Sarah De Coster impressed me with her interests on weight lifting.
- Sarah Megens is always a generous friend with passions on HIV research.
- Soo-Yon Rhee enlightened me on future research.
- Soraya Maria Menezes consistently supports me since the very beginning of my Phd.
- Steven Sijmons shared with me many working nights in the lab.
- Stijn Imbrechts taught me his magic skills on software development.
- Supinya Piampongsant gave me valuable help on my publications.
- Susan Obeid has been a good friend.
- Thomas Mina has been a nice friend shared many beers in oude market.
- Tiago Graf shared with me the exciting thrill of cobra coaster.
- Tim Dierckx has been hard working besides my desk in lab.
- Valentijn Vergote has been a good friend sharing many ideas.
- Yoeri Schrooten always organized pleasant and memorable parties.

I would like to acknowledge my flatmates who have brought laugh and peace during my life in Leuven. Especially, I would like to mention family members who have lived at Schapenstraat 33: Hernán Blanco Landa, David Martinez Barreiro, Dan Clements, Giwa Babatunde, Simona Dobos, Margarita Buliy, Kamuti Mulonda, Daniele Strafile, Julio Villalvazo, Sultan Salahuddin, Yiteng Hou, Merijn De Coster, Ankit Mehta, Carlos Fernández Reyes, Meng Zhang, Natalia Parkhacheva and Natalia Andreeva.

I would like to thank my classmates during my course studies: Amin Davani, Gloria Fabris, Chen Jialin, Justyna Startek, Ying Ting, Chen Ziwei and Kevin Louris. It is because of you, boring days sitting in the quiet classrooms became enjoyable memory.

Lastly, I would like to say a few words to my family. As a son, I am indebted to my parents. As a brother, I owe much to my sister. Their selfless sacrifice and consistent support are the reasons I can present to you this Phd thesis.

Table of contents

Abbreviations

3TC	Lamivudine	**MA**	Matrix
AA	Amino acid	**MAG**	Maximal ancestral graph
ABC	Abacavir	**MBC**	Multi-dimensional Bayesian classifier
AG	Ancestral graph	**MCMC**	Markov chain Monte Carlo
AIDS	Acquired immune deficiency syndrome	**MHC**	Major histocompatibility complex
AP	Ancestral polytree	**MI**	Mutual information
APC	Average product correction	**MIBP**	Mutual information biochemical property
ART	Antiretroviral therapy	**MPER**	Membrane-proximal external region
ASA	Solvent accessible surface area	**MSA**	Multiple sequence alignment
ASC	Average sum correction	**NC**	Nucleocapsid
ATV	Atazanavir	**NCPS**	Normalized coevolutionary pattern similarity
AUC	Area under the precision-recall curve	**Nef**	Negative regulatory factor
BDS	Bayesian Dirichlet score	**NFV**	Nelfinavir
BN	Bayesian network	**NHR**	N-terminal heptad repeat
BSC	Bregman soft clustering	**NIH**	National Institutes of Health
CA	Capsid	**NN**	Neural network
CASP	Critical assessment of protein structure prediction	**NNRTI**	Non-nucleoside analog reverse-transcriptase inhibitor
CHR	C-terminal heptad repeat	**NPC**	Nuclear pore complex
CI	Conservation index (chapter 2)	**NP-hard**	Non-deterministic polynomial-time hard
CI	Conditional independence (chapter 7)	**NRTI**	Nucleoside analog reverse transcriptase inhibitor
CI	Confidence interval (chapter 3,4,5,6)	**NTD**	N-terminal domain
CL	Cytoplasmic tail	**NVP**	Nevirapine
CNPR	Optimized method combination	**OPAP**	Orienting principles of ancestral polytree
CPS	Coevolutionary pattern similarity	**OR**	Odds-ratio
CRF	Circulating recombinant form	**ORF**	Open reading frame
CSM	Cleavage site mutation	**p1**	Spacer peptide 1
CTD	C terminal domain	**p2**	Spacer peptide 2
CTMP	Continuous time Markov process	**p6***	Transframe peptide in the gagpol precursor
d4T	Stavudine	**PBMC**	Peripheral blood mononuclear cell
DAG	Directed acyclic graph	**PCC**	Pearson's correlation coefficient
DCA	Direct coupling analysis	**PDB**	Protein data bank
DDDP	DNA-dependent DNA polymerase	**PDFS**	Polytree depth first search
ddI	Didanosine	**PI**	Protease inhibitor
DGB	Drug genetic barrier	**PIC**	Pre-integration complex
DLV	Delavirdine	**PPI**	Protein-protein interaction
DN	Deep network	**PR**	Protease
DNA	Deoxyribonucleic acid	**PRA**	Polytree recovery algorithm

DRV	Darunavir	**RC**	Replication capacity
dsDNA	Double-stranded DNA	**RCW**	Row and column weighting
EC50	50% effective concentration	**RDDP**	RNA-dependent DNA polymerase
ECS	Ensemble coevolution system	**Rev**	Regulator of virion expression
EFV	Efavirenz	**RMSE**	Root-mean-square error
EI	Entry inhibitor	**RNA**	Ribonucleic acid
EM	Expectation maximization	**RNN**	Recursive neural network
env	Envelope gene	**RRE**	Rev response element
ER	Endoplasmic reticulum	**RT**	Reverse transcriptase
ETR	Etravirine	**RTC**	Reverse transcriptase complex
FCI	Fast causal inference algorithm	**RTV**	Ritonavir
FDA	U.S. Food and Drug Administration	**SAP**	Simple ancestral polytree
FDR	False discovery rate	**SCA**	Statistical coupling analysis
FPV	Fosamprenavir	**SIV**	Simian immunodeficiency virus
FTC	Emtricitabine	**SLAC**	Single likelihood ancestor counting model
gag	Group-specific antigen gene	**SQV**	Saquinavir
GCS	Gag cleavage site	**ssDNA**	Single-stranded DNA
GP41CT	Cytoplasmic tail of GP41	**SVM**	Support vector machine
GST	Glutathione-S-transferse	**TAR**	Transactivating response element
GTR	Generalized time-reversible model	**Tat**	Trans-activator of transcription
HAART	Highly active antiretroviral therapy	**TDF**	Tenofovir
HBV	Hepatitis B virus	**TM**	Transmembrane domain
HCV	Hepatitis C virus	**TPV**	Tipranavir
HIV	Human immunodeficiency virus	**UTR**	Untranslated region
HLA	Human leukocyte antigen	**Vif**	Viral infectivity factor
IC50	50% inhibitory concentration	**VMD**	Visual molecular dynamics
IDV	Indinavir	**Vpr**	Viral protein R
IN	Integrase	**Vpu**	Viral protein U
LAP	Learning ancestral polytree	**WHO**	World health organization
LPV	Lopinavir	**XCS**	Extended classifier system
LTR	Long terminal repeat	**ZRES**	Z-residue score

Summary

Human Immunodeficiency Virus (HIV) has been a worldwide threat to public health and the economy during the last three decades. About 35.3 million people were living with HIV at the end of 2012. Currently, no effective vaccine or cure is available despite the development of multiple antiretroviral drugs for HIV treatment.

Extensive genetic variation and rapid evolution has been observed in the HIV genome, making HIV one of the fastest evolving organisms and particularly, challenging the development of drugs and vaccines. Because of this, drug resistance mutations emerging in the HIV genome can cause failure of antiviral therapy for all drug classes. The HIV genome only encodes fifteen viral proteins. During the HIV life cycle, fast viral replication and infection require a high level of interactions between HIV proteins, as well as interactions between HIV and human proteins. Therefore, a comprehensive analysis of HIV genomic diversity, interaction and coevolution can provide insights on the development of new drug classes and vaccines.

The objectives of this Phd study were to investigate HIV genome-wide diversity, interaction and coevolution. We focused on the development of computational methods, which can be used for genome-wide analyses on large-scale datasets of genomic sequences, protein structures, anti-HIV inhibitors, HIV-human protein interactions and human immunological data. The structure of this thesis is organized as follows.

Chapter 1 introduces the background knowledge about HIV genome, evolution, life cycle, antiretroviral treatment and functional cure. Information of HIV-1 genome-wide inter-protein interactions is further reviewed.

Chapter 2 presents the first study to evaluate the functional conservation of HIV-1 Gag proteins and to identify natural variations at the drug binding sites of Gag inhibitors. Using more than 10000 viral sequences, this study highlights natural variations of known drug binding sites and identifies the conserved drug targets in the HIV-1 Gag proteins.

Chapter 3 extends the idea from Chapter 2 and presents the first study to characterize HIV genomic diversity using large-scale genomic datasets. We integrated data accumulated during the last three decades including approximately 3000 full-length genome sequences, all HIV protein structures, detailed HIV-human protein interaction data, human immunological data and all peptide inhibitors derived from the HIV genome. Using this large-scale data, we measured genomic diversity of major HIV clades, determined factors shaping HIV genomic diversity, identified potential HIV vaccine strains, demonstrated conserved drug-target regions and characterized known peptide inhibitors derived from HIV-1 full-length genome.

Chapter 4 evaluates HIV-1 Gag mutations emerging under drug selective pressure, for which we aimed to understand the impact of genome-wide coevolution in HIV drug resistance. We identified HIV-1 Gag mutations that were significantly associated with genotypic drug resistance to protease inhibitors (PIs). This study reports the first large-scale analysis to evaluate PI-associated Gag mutations emerging during the PI treatment.

Chapter 5 introduces a new ensemble coevolution system designed for predicting the intra- and inter-protein coevolution in the HIV-1 genome. This system integrates 27 sequence-based methods published in the past decade. We designed a heuristic algorithm to identify a combination of four methods that outperformed any individual method in predicting HIV-1 intra- and inter-protein coevolution.

Chapter 6 applies our ensemble coevolution system with the method combination optimized in Chapter 5 to model the HIV-1 Gag-protease coevolution networks. We showed that HIV-1 cleavage site mutations and Gag C-terminal mutations coevolving with protease drug resistance mutations. These PI-associated Gag positions may interact with human proteins, but they may not affect Gag inhibitors because drug binding sites are unlikely to coevolve with protease positions.

Chapter 7 proposes ancestral polytrees as new graphical models to investigate large-scale interaction networks. We showed that ancestral polytrees were efficient to model mutation pathways in HIV-1 proteins.

Chapter 8 discusses the ideas of my projects, as well as the strength and weakness of my studies. Data visualization and future perspectives of my projects are briefly discussed.

Overall, this thesis contributes to the understanding of HIV genome-wide diversity, interaction and coevolution.

Samenvatting

Het Humaan Immunodeficiëntie Virus (HIV) is in de afgelopen 30 jaar uitgegroeid tot een wereldwijde bedreiging voor de wereldgezondheid en -economie, zo waren er op het einde van 2012 35.3 miljoen mensen besmet met HIV. Ondanks uitgebreid wetenschappelijk onderzoek naar HIV in de afgelopen 30 jaar.is er op dit moment nog altijd geen vaccinatie of genezing mogelijk. Wel zijn er verschillende antiretrovirale medicijnen ontwikkeld die de progressie naar het ziektestadium van AIDS vertragen door de virale replicatie te blokkeren op verschillende stages van de virale levenscyclus. Behandeling van HIV infectie is echter nog levenslang noodzakelijk.

Het HIV genoom wordt gekenmerkt door een hoge mate van genetische variabiliteit en snelle evolutie, wat ervoor zorgt dat HIV als één van de snelst evoluerende organismen is. Deze evolutionaire flexibiliteit zorgt er op zijn beurt voor dat het moeilijk is om een algemeen werkend vaccin of medicijn te ontwikkelen. Mutaties in het virale genoom kunnen leiden tot de ontwikkeling van virale resistentie tegen antiretrovirale geneesmiddelen en uiteindelijk tot het falen van de behandeling. Het virale genoom codeert voor 15 verschillende virale eiwitten, die de nodige interacties aangaan met zowel andere virale eiwitten als met eiwitten van de gastheercel om de replicatie van het virus en de effectieve infectie van gastheercellen te bewerkstelligen. Het is dus van belang om een beter inzicht te verkrijgen in de bestaande genetische diversiteit van HIV en de evolutionaire dynamiek van eiwit interacties, zodat de verkregen kennis kan bijdragen tot de ontwikkeling van nieuwe antiretrovirale geneesmiddelen en vaccins tegen HIV.

Het doel van deze doctoraatsstudie was dan ook het in kaart brengen van diversiteit in het HIV genoom en het modeleren van co-evolutie. Er werd een sterke nadruk gelegd op de ontwikkeling en toepassing van computationele methoden die geschikt zijn voor analyses van grote datasets met informatie over HIV genoom sequenties, eiwitstructuren, geneesmiddelen, interacties tussen verschillende virale en humane eiwitten alsook beschikbare klinische en immunologische data van HIV patiënten.

In het eerste inleidende hoofdstuk worden de verschillende algemene facetten van HIV geïllustreerd. In het kort worden de genoomstructuur, evolutie en levenscyclus van HIV beschreven evenals de aspecten rond bestaande preventiemaatregelen, antiretrovirale behandeling en potentiële genezing van HIV infectie.

Het tweede hoofdstuk beschrijft een studie naar de natuurlijke variatie van HIV-1 Gag eiwitten en in het bijzonder van eiwit posities die belangrijk zijn voor de binding met HIV-1 antiretrovirale geneesmiddelen. Door gebruik te maken van meer dan 10000 virale sequenties werd amino zuur diversiteit in gekende interactie- en bindingsdomeinen

voor verschillende Gag inhibitoren in het virale Gag eiwit en in mogelijke nieuwe doelwitten voor therapie in kaart gebracht.

Het derde hoofdstuk vertrekt van een gelijkaardig idee als het tweede hoofdstuk maar presenteert de eerste studie die de genetische diversiteit van het hele HIV genoom bestudeert door gebruik te maken van grootschalige datasets. In de afgelopen 30 jaar is er een aanzienlijke hoeveelheid aan informatie verzameld over verschillende aspecten van HIV infectie. In deze studie hebben we verschillende bestaande datasets met elkaar geïntegreerd, met betrekking tot 3000 volledige HIV genoom sequenties, alle HIV eiwit structuren, gedetailleerde informatie over verschillende HIV-gastheer eiwit-interacties, klinische en immunologische data van HIV-patiënten en informatie over alle HIV peptide inhibitors die ontwikkeld zijn gebaseerd op het HIV genoom. Op basis van deze informatie hebben we de genetische diversiteit in het HIV genoom nauwkeurig beschreven, factoren geïdentificeerd die deze genetische diversiteit vorm geven, HIV stammen voorgesteld voor de ontwikkeling van een HIV vaccin en tot slot gekende HIV peptide inhibitoren, ontwikkeld op basis van het HIV-1 genoom, gekarakteriseerd.

Hoofdstuk vier evalHIV-1 Gag mutaties die worden geselecteerd in verschillende HIV-1 subtypen onder selectieve druk van bestaande antiretrovirale medicijnen. Verschillende mutaties in Gag werden geïdentificeerd die significant geassocieerd waren met genotypische drug resistentie tegen protease inhibitoren (Pis). Deze studie rapporteert de eerste grootschalige analyse die PI-geassocieerde mutaties bestudeert die ontstaan tijdens de behandeling met PIs.

Het vijfde hoofdstuk introduceert en beschrijft de ontwikkeling van een nieuw ensemble model dat toelaat om intra- en inter-eiwit co-evolutie in het HIV-1 genoom te analyseren. Dit systeem integreert 27 bestaande sequentie-gebaseerde predictie modellen, die gepubliceerd zijn in het afgelopen decennium. Er wordt vervolgens aangetoond dat dit samengestelde systeem een optimale combinatie van methoden identificeerde dat betere resultaten opleverde dan de meeste individuele modellen in het voorspellen van HIV-1 intra- en inter-eiwit co-evolutie.

Hoofdstuk zes past de optimale combinatie van methoden, geïdentificeerd in het ensemble model van het vijfde hoofdstuk, toe om HIV-1 Gag-protease co-evolutie netwerken in HIV-1 subtype B te bestuderen. Er wordt aangetoond dat mutaties in het C-terminale domein van Gag of in Gag klieving posities co-evolueren met verschillende protease resistentie mutaties. Bovendien wordt er aangetoond dat deze Gag mutaties mogelijks kunnen interageren met verschillende humane eiwitten, maar niet met e bestaande Gag inhibitoren, omdat de drug bindingsdomeinen in het Gag eiwit over het algemeen niet mee evolueerde met de verschillende mutaties in het protease enzyme..

Het zevende hoofdstuk presenteert verschillende "ancestral polytree networks" als potentiële nieuwe probabilistische grafische modellen voor het onderzoeken van grootschalige interactie netwerken op eiwit niveau. Er wordt aangetoond dat deze modellen efficiënt de verbanden tussen mutaties in verschillende HIV-1 eiwitten kunnen modelleren.

Kort samengevat, hebben we in dit proefschrift HIV-1 genoom diversiteit, de associatie met therapie selectieve druk en factoren die een impact hebben op diversiteit onderzocht. Verder werd er ook een systeem op punt gesteld om HIV-1 co-evolutie te modeleren. Een ideale combinatie van individuele methoden werd geïdentificeerd die een betere predictie opleverde dan elk van de methoden afzonderlijk.

Chapter 1

General introduction

"To err is human, to forgive, divine."

— Alexander Pope

1.1 Human immunodeficiency virus

In 1981, a new disease recognized as Acquired Immune Deficiency Syndrome (AIDS) was spreading among young homosexual men in America, who were succumbing to uncommon opportunistic infections and rare malignancies [1]. In 1983, the discovery of a retrovirus, now termed human immunodeficiency virus (HIV), was recognized as the causative agent of AIDS [2]. Since then, HIV/AIDS has become one of the most devastating infectious diseases that have emerged in recent history. At the end of 2012, the WHO global health report showed that about 35.3 million people were living with HIV (http://www.who.int/gho/hiv/en/). Over 12 million children have been orphaned by AIDS and about 1600 babies acquire HIV from their infected mothers every day [1]. In the past three decades, the HIV pandemic has caused a great burden to the global wealth and health, especially in Sub-Saharan Africa where the highest rate of HIV infection has been recorded (http://www.unaids.org/). As the prospects of effective vaccine and curative treatments remain uncertain, HIV/AIDS will continue to be a significant threat to public health in the coming years.

Based on genetic similarities, several HIV lineages have been identified including HIV type 1 (HIV-1) groups M, N, O, P and HIV type 2 (HIV-2) groups A-H. HIV-1 group M causes the majority (>90%) of HIV global infections. HIV-1 group O has

infected a few tens of thousands of patients in West-Central Africa, HIV-1 group N has only been found in a small number of people in Cameroon and HIV-1 group P was recently reported in two patients originating from Yaounde, Cameroon [3]. Nine subtypes (A-D, F-H, J, K) have been classified in the HIV-1 group M. As a major subtype, HIV-1 subtype C accounts for nearly half (48%) of the HIV-1 global infections, while subtype B dominates infections in Europe and America. Besides HIV-1 subtypes, the recombination between different HIV-1 subtypes has generated more than 50 circulating recombinant forms (CRFs), contributing to the global HIV diversity [4].

HIV originated from multiple zoonotic transmissions from non-human primates (chimpanzee, western gorilla, sooty mangabey) to humans in West-Central Africa [3]. These zoonotic transmissions probably happened during the hunting and butchering of primates for bushmeat, as well as the capture, trade and keeping of monkeys as pets [5]. Moreover, simian immunodeficiency virus (SIV) has been identified as the ancestor of different HIV lineages in humans [3]. For instance, HIV-1 groups M and N originated independently from SIVcpz in chimpanzees living in West-Central Africa. HIV-1 groups O and P originated from SIVgor identified in Western lowland gorillas living in Cameroon. Notably, divergent SIVcpz lineages are the ancestors of SIVgor strains [3]. Accumulated evidence has shown that Kinshasa in the Democratic Republic of the Congo might be the birth of the global HIV-1 epidemic [6].

The estimated date of HIV origins has been traced back to the late 19^{th} or the early 20^{th} century when many factors exerted a strong influence during the course of HIV transmission [3, 5]. The earliest direct evidence of HIV infection was identified retrospectively in a serum sample and a lymph node biopsy specimen collected from Kinshasa in 1959 and 1960, respectively [7]. The date for the cross-species transmission between HIV-1 group M and SIVcpz was estimated to be 1853 (95% confidence interval: 1799-1904). The estimated date of the group O origin was about 1920 (1890-1940) and the transmission of HIV-1 group N probably took place around 1921 (1885-1955) [3]. The transmission time of major HIV-1 subtypes and CRFs was largely after 1950s [8].

1.2 HIV genome

Fifteen viral proteins are coded by nine genes in the HIV genome (Figure 1.1). Three major genes, *gag*, *pol* and *env*, code for structural proteins (Matrix, Capsid, Nucleocapsid, p6), viral enzymes (Protease, Reverse transcriptase, Integrase) and envelope proteins (GP120, GP41). The remaining genes code for regulatory proteins (Tat, Rev) and accessory (so-called auxiliary) proteins (Vif, Vpr, Vpu/Vpx, Nef) [9]. Although both HIV-1 and HIV-2 originated from SIV [3], they have different gene maps (Figure 1.1). Particularly, Vpu in HIV-1 and Vpx in HIV-2 mark a distinct difference between these two HIV types.

Interestingly, 11 of 16 HIV proteins are multimeric proteins folded with multiple protein units (chains) such as Matrix [10], Capsid [11], Protease [12], RT [13], integrase [14], Vif [15], Tat [16], Rev [17], GP120-GP41 [18] and Nef [19] (Figure 1.1). Multimeric HIV proteins play crucial roles during the viral life cycle. For instance, Matrix and capsid multimers are needed to construct the viral structures of mature virions [10, 11]. Protease dimers catalyze the Gag and GagPol polyproteins during viral maturation [12]. Reverse transcriptase dimer is required to produce viral dsDNA during reverse transcription [13]. Integrase tetramer is essential for integrating viral dsDNA into host chromosomes during viral integration [14]. Rev multimers exports viral mRNA from nucleus to cytoplasm [17]. Envelope trimers formed with GP120 and GP41 interact with cellular receptors (e.g. CD4) during viral entry [18]. Nef dimers are required for the downregulation of receptors on plasma membrane surface [19]. We briefly summarize the HIV protein functions.

Matrix: Matrix is a structural protein encoded by the *gag* gene, which provides the basic infrastructure of HIV particles. The matrix domain in the intact Gag polyprotein is destined to traffic Gag to the plasma membrane for viral budding and to recruit host factors (e.g. TIP47) [20]. Matrix anchors the lipid membrane through the myristoylated N-terminal domain which is critical for plasma membrane targeting and viral assembly [20]. When cleaved from Gag polyprotein during viral maturation, Matrix trimmers organize into ordered hexamers to build a structural layer beneath viral membrane, which protects the integrity of HIV particles [21]. To prevent the

nonspecific binding, matrix in the Gag polyprotein binds nucleic acids through a PIP2-dependent mechanism [21].

Capsid: Capsid is a structural protein encoded by the *gag* gene which provides the basic infrastructure of viral particles [21]. The hexamer and pentamer structure of Capsid constitutes the conical fullerene core of mature HIV particles [22]. The interaction between Capsid and host proteins allows for the packaging of host proteins (e.g. cyclophilin A) into HIV particles. Capsid also binds with the host restriction factor TRIM5α to prevent viral uncoating at the early stage [23].

Nucleocapsid: Nucleocapsid is a structural protein encoded by the *gag* gene [21, 24-26]. To prevent viral RNA from nucleases, Nucleocapsid binds with the genomic viral RNA during viral packaging and coats the genomic RNA within viral core [27]. Nucleocapsid can also bind to host proteins such as the ESCRT-associated protein ALIX to promote viral budding [28]. Served as an RNA chaperone, nucleocapsid enhances nucleic acid-dependent steps in the HIV life cycle. For instance, it promotes the DNA strand exchange reactions during reverse transcription and stimulates viral integration during viral integration [29].

p6: p6 is a structural protein at the C terminus of *gag* gene [21]. p6 can recruit the host machinery to bud the virus outwards from the cell surface [30]. Viral protein Vpr and host proteins (e.g. AIP1/ALIX) bind to p6 during viral packaging [21].

Protease: The first viral enzyme encoded by the *pol* gene is protease. During viral maturation, protease cleaves Gag polyproteins at the cleavage sites to produce structural proteins (Matrix, Capsid, Nucleocapsid, p6). Protease cleaves the GagPol polyproteins to produce viral enzymes (Protease, Reverse transcriptase, Integrase). Moreover, the activity of protease depends on the concentration of GagPol polyproteins and the rate of protease-mediated autoprocessing is modulated by the adjacent p6 sequence [31].

Reverse transcriptase (RT): RT is another important enzyme encoded by the *pol* gene. To produce dsDNA from the viral single-stranded RNA genome, RT in the reverse transcriptase complex (RTC) catalyzes both the RNA-dependent and the DNA-dependent DNA polymerization reactions. During reverse transcription, RT

jumps from one template to another when two copies of single-stranded genomic RNAs exist per virion. The frequent template switch promotes the generation of novel recombinant DNA genome sequences derived from two parental RNA sequences [29]. Many mutations occur because HIV reverse transcription is highly error-prone.

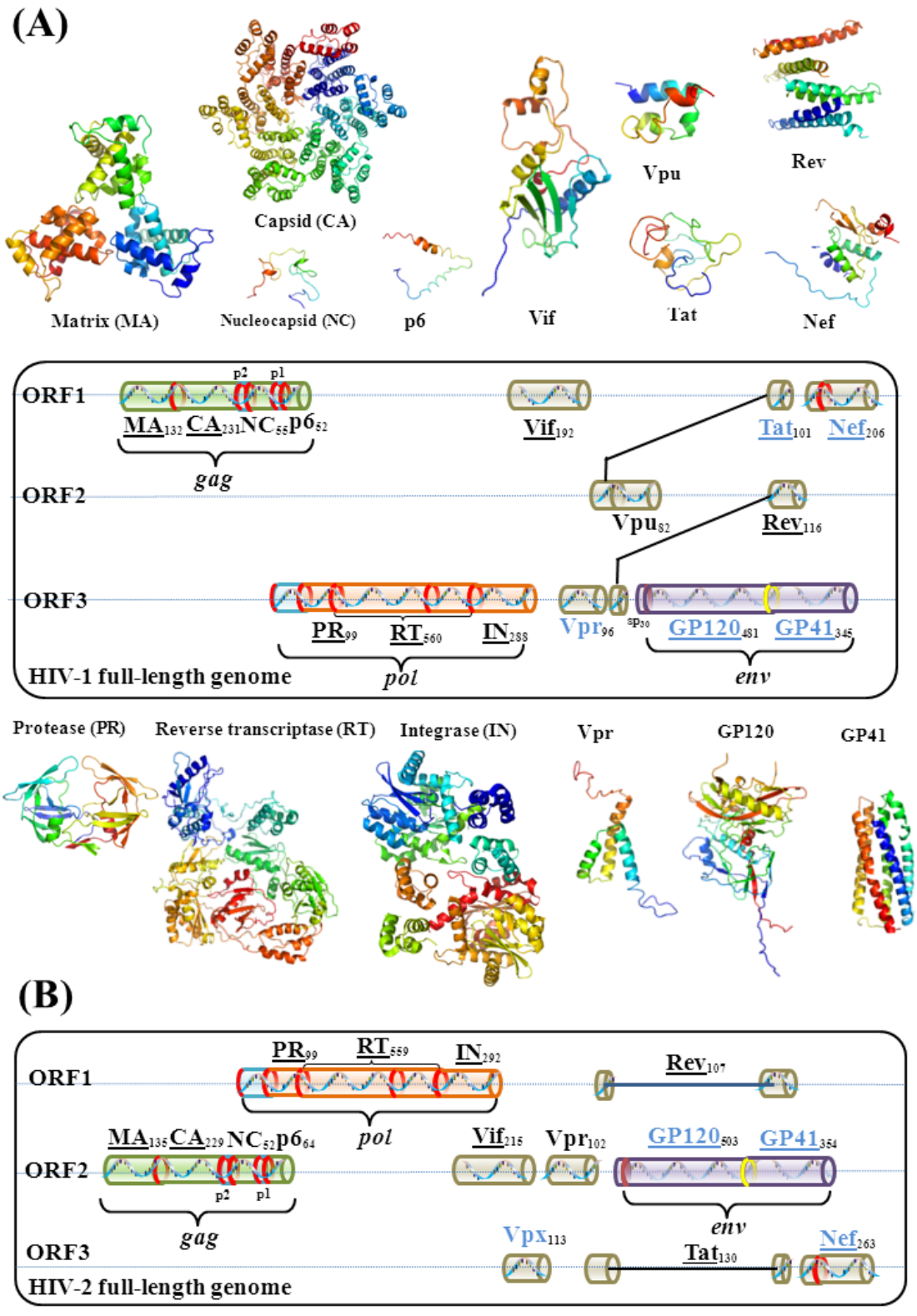


Figure 1.1: Gene map and protein structure of HIV-1 and HIV-2 at the full-length genome. Protein names on the gene map are annotated with their amino acid lengths (HIV-1 reference: HXB2, HIV-2 reference: BEN). HIV multimeric proteins are indicated by underlined names and extracellular proteins have their names colored blue. Red rings in the gene map indicate the locations where HIV-1 protease cleaves during viral maturation. The yellow ring in the *env* gene indicates the site for human proteases (furin, PC1) to cleave [32]. Protein domains indicated in the gene map are not to scale. Abbreviations: MA: matrix, CA: capsid, p2: spacer peptide 2, NC: nucleocapsid, p1: spacer peptide 1, PR: protease, RT: reverse transcriptase, IN: integrase, Vif: viral infectivity factor, Vpr: viral protein R, Tat: trans-activator of transcription, Vpu: viral protein U, Rev: regulator of virion expression, sp: signal peptide, GP120: surface glycoprotein GP120, GP41: transmembrane glycoprotein GP41, Nef: negative regulatory factor, *gag*: group-specific antigen gene, *env*: envelope gene, ORF: open reading frame.

Integrase: Integrase is the third enzyme encoded by the *pol* gene. After the nuclear import of pre-integration complex (PIC), viral Integrase catalyzes two major reactions (3'-processing and strand transfer reactions) to insert the linear, double-stranded viral DNA into human chromosomes. In the mature viral particles, Integrase is cleaved from the GagPol polyprotein by viral Protease. Moreover, reverse transcriptase binds with integrase to prevent the catalytic activity of Integrase before viral integration [33]. As part of reverse transcriptase complex, Integrase also plays a role during reverse transcription [29].

GP120: Encoded by the *env* gene, the surface glycoprotein GP120 is exposed on the surface of HIV particles [34]. On the virion surface, there are less than 30 envelope spikes consisting of three molecules of GP120 and GP41 each, connected by non-covalent interactions [29]. During viral entry, GP120 interacts with specific receptors (e.g. CD4) on cell surface [35]. Specifically, the binding of CD4 to the third and fourth loop regions of GP120 induces the conformational changes of GP120, which exposes the V3 loop of GP120 to interact with cellular coreceptors (e.g. CCR5). Many human neutralizing antibodies have been found to target GP120 in a strain-specific manner, while a few antibodies (e.g. PG9, PG16) have a broad neutralization activity against different HIV-1 strains [36-38].

GP41: The transmembrane glycoprotein GP41 is also encoded by the *env* gene. GP41 contains a glycine-rich region which is essential for the membrane fusion activity [39]. Multiple functions of GP41 have been reported [39]: (1) the intracellular trafficking of

the Env protein is regulated by the cytoplasmic tail of GP41 (GP41CT) which interacts with various cellular proteins. (2) GP41CT interacts with the viral Matrix protein to regulate the Env incorporation into HIV virions. (3) GP41CT regulates the internalization exerted by the clathrin-mediated endocytosis. (4) GP41CT regulates cellular activation of host transcription factors (e.g. NF-kB). (5) GP41 interacts with host proteins to regulate the activity of actin cytoskeleton. (6) HIV-1 GP41 membrane-proximal external region is targeted by human antibodies (e.g. 10E8) with a broad neutralization activity [40].

Vif: Viral infectivity factor is an accessory protein encoded by all lentiviruses except the equine infectious anemia virus [41]. Vif is famous to hijack the human ubiquitin ligase complex CBF-β to counteract the antiviral activity of host proteins, APOBEC-3G and APOBEC-3F, both of which interfere with the correct assembly of HIV-1 viral core [42, 43]. Vif also interacts with Gag polyprotein to modulate the Protease-mediated proteolytic processing [41]. Vif is incorporated in HIV particles [41].

Vpr: Viral protein R is an accessory protein which plays multiple functions to enhance HIV replications in the non-dividing cells (e.g. macrophages). Vpr plays multiple functions such as the modulation of viral reverse transcription, the nuclear import of HIV-1 pre-integration complex, the transactivation of HIV-1 long terminal repeat (LTR) promoter, the induction of apoptosis and G2 cell cycle arrest (see review [44]). Vpr is incorporated in HIV particles [45].

Vpu: Viral protein U is a membrane-associated accessory protein with two major functions (CD4 downregulation, Tetherin antagonism) [46]. First, Vpu hijacks the human ubiquitin machinery to target CD4 and induces the downregulation of CD4 receptors in the endoplasmic reticulum (ER). Second, Vpu antagonizes Tetherin, an interferon-regulated human restriction factor, to enhance the release of viral particles in a cell-type dependent manner. Vpu is not incorporated in HIV particles [47].

Vpx: Vpx is an accessory protein in HIV-2 but absent in HIV-1. Major functions of Vpx include: (1) Vpx induces the ubiquitin-proteasome-dependent degradation of SAMHD1 [48-50], which restricts the HIV-2 replication in myeloid cells. (2) Vpx is required for HIV-2 reverse transcription [51]. (3) Vpx assists the nuclear import of viral pre-integration complex (PIC) [48-50]. A major difference between Vpr in HIV-

1 and Vpx in HIV-2 is that Vpr arrests the host cell cycle in the G2 phase, while Vpx targets the host restriction factor SAMHD1 for the proteasomal degradation [48]. Despite these, common attributes have been reported [49, 50]: (1) both are originated from the same ancestral gene; (2) both are incorporated into HIV particles via the interaction with p6 in *gag* precursors; (3) both are involved in the nuclear import of pre-integration complex. Vpx is not incorporated in viral particles.

Rev: Rev is an accessory protein which controls the nuclear export of unspliced and partially spliced viral RNAs from nucleus to cytoplasm [52]. Rev multimers bind to the stem-loop structure of Rev response element (RRE) in the *env* coding region of viral RNA, forming a large oligomeric ribonucleoprotein (RNP) [29]. RNP complex interacts with the human export factor CRM1 (exportin 1 or Xpo1) and shuttles through the nuclear pore complex (NPC) from nucleus to cytoplasm. Overall, Rev activity exerts a strong influence on HIV-1 RNA transport, translation and packaging [53]. Rev is not incorporated in the viral particles [47].

Tat: Trans-activator of transcription is a regulatory protein which plays essential roles in viral replication. Tat exists in all lentiviruses and is the first eukaryotic transcription factor known to interact with TAR (transactivating response element) in RNA instead of DNA [29]. Tat interacts with many human proteins to execute multiple functions [29, 54, 55]: (1) Tat activates the transcription initiation and elongation of HIV-1 LTR promoter, preventing the premature termination of transcription and polyadenylation. (2) Tat acts as a nucleic acid chaperone to regulate the capping of HIV-1 mRNA. (3) Tat induces the T cell apoptosis, neurodegeneration and oxidative stress. (4) Tat regulates the expression of major histocompatibility complex (MHC) and downregulates several cell surface receptors. (5) Tat suppresses the activity of reverse transcriptase to prevent the premature synthesis of viral DNA. (6) Extracellular Tat upregulates the CXCR4 expression on CD4+ T cells, stimulates the expression of cytokines and interacts with cell-surface receptors to activate cellular signal transduction pathways. Tat is not incorporated in the viral particles.

Nef: Negative regulatory factor is an accessory protein which enhances viral pathogenesis [56]. During the viral life cycle, Nef can play multiple roles [56]: (1) Nef downregulates CD4 receptors and MHC molecules; (2) Nef promotes the viral release and the cell-to-cell transmission; (3) Nef activates the apoptosis and involves

with the clathrin-dependent endocytic pathways. Nef is incorporated in the viral particles [56].

1.3 HIV evolution

Extensive genetic variation has been observed at the HIV genome, making HIV one of the fastest evolving organisms [57]. Rapid HIV evolution is the result of multiple factors: (1) a high substitution rate (~0.2 errors per genome during each replication cycle) [57], (2) a high replication rate (~10^{10}-10^{12} new particles per day) [57], (3) a high recombination rate ($1.4 \pm 0.6 \times 10^{-5}$ recombinations per site and generation)[58]. Because of these, HIV is famous for a high genomic diversity, which is associated with disease progression [59]. In fact, HIV genetic diversity of plasma isolates is reduced at any time point but increases during the course of infection [59]. Different processes have been reported to drive the HIV genetic diversity [57, 59]. For instance, The bottleneck accompanying HIV transmission can greatly reduce the genetic diversity [57], whereas natural selection is a less potent force to increase genetic diversity among hosts than within hosts [59]. Rapid HIV evolution generates T cell escape mutants to elude host cytotoxic T lymphocyte responses, causing a major barrier to develop effective HIV vaccines [57].

Recombination: Recombination plays a central role in generating HIV genetic diversity during viral evolution. Genetic recombination occurs when the HIV reverse transcriptase switches between alternative genomic templates during reverse transcription. Moreover, recombination has been reported in all phylogenetic levels: among primate lentiviruses, among HIV-1 groups, among subtypes and within subtypes [57]. Circulating recombinant forms (CRFs) are the inter-subtype recombinants fixed in the patient populations. Until now, over 50 CRFs have been reported including the most prevalent CRF01_AE and CRF02_AG in the HIV epidemic [4]. Recombination is associated with natural selection and genetic drift to generate complex population dynamics, providing an efficient mechanism for HIV to escape the accumulation of deleterious mutations [60].

HIV reservoir: HIV reservoirs serve to maintain the pool of replicating virus in various cell types such as CD4+ T lymphocytes, follicular dendritic cells and macrophages [57]. HIV reservoirs protect HIV from anti-HIV inhibitors and promote

greater genetic diversity than non-reservoir virus due to the reservoir of archival strains [61]. Antiretroviral treatment can impair viral replication but has limited efficiency to eliminate HIV reservoirs [62]. It is partly because immune privileged tissues can drive anatomical compartments to act as sanctuary sites where HIV can still replicate beyond the reach of ant-HIV inhibitors [61]. The identified sanctuaries include the central nervous system, the gut-associated lymphoid tissue and the genitourinary tract [63]. Potential approaches have been proposed to eliminate HIV reservoirs. Most strategies are based on the principle of activating cells with HIV reservoir to induce viral expression from the HIV genome; for instance, the stem cell immune reconstitution and the activation-elimination strategy. Unfortunately, none of the published strategies has been proven to purge all latent virus complexity[62].

HIV evolution differs in different human tissues. HIV evolution has shown to be different in specific human tissues [64-68]. For instance, the HIV evolutionary rate can be differed between different brain compartments [65]. The higher evolutionary rate in the meninges and temporal lobe can be caused by the enhanced infection rate of macrophages during immune system failure [65]. Independent evolution of macrophage-tropism and increased charge in HIV-1 envelope proteins has also been found between the human brain and immune tissues [66]. Other tissues such as choroid plexus, bone marrow, lung and liver also provide different reservoirs for HIV pathogenesis [67].

Due to high replication and mutation rates, HIV accumulates enormous mutations in its viral quasispecies - a group of viruses associated by similar mutations. By generating the viral quasispecies, HIV can induce the emergence of drug resistance mutations to cause the failure of HIV antiviral therapy. Overall, a better understanding of HIV evolution and diversity can benefit current antiviral treatment, as well as the drug and vaccine design. To efficiently eradicate HIV reservoirs remains a challenge in the coming years.

1.4 HIV life cycle

In a therapy-naïve patient, HIV-1 produces about 10 billion nascent virions and infects 100 million cells per day [69]. Such fast replication is ensured by the HIV life

cycle which can be divided into six major stages: viral entry, reverse transcription, integration, transcription/translation, budding and maturation (**Figure 1.2**).

Viral entry is the process of mature virions entering host cells, after which reverse transcriptase produces double-stranded DNA (dsDNA) from the viral genomic RNA. dsDNA is then imported into nucleus where viral integrase integrates the dsDNA into host chromosomes. Transcription makes large amounts of viral RNA copies from the integrated viral DNA. The export of viral RNAs from nucleus to cytoplasm allows for the translation of folded (precursor) proteins. During viral budding, viral genomic RNA and the Vif, Nef, Vpr, Gag, GagPol and Env proteins are assembled to build nascent HIV particles. During viral maturation, Gag and GagPol polyproteins are cleaved by viral Protease. Besides these major stages, the endocytosis, exocytosis and cell-to-cell transmission provide alternative pathways for viral entry and budding under certain conditions [70].

HIV-human protein interactions play essential roles for HIV to hijack human cellular systems for viral replication [71-73]. For instance, Vif initiates the ubiquitination and degradation of the human protein APOBEC3G via the proteasomal pathway. Previous studies have investigated the global landscape of HIV–human protein interactions using the genome-wide siRNA libraries [74] and the quantitative scoring system MiST [75]. Our knowledge on HIV-human protein have also been enriched by the established databases such as the HIV-1 human protein interaction database [76] and the VirusMINT database [77].

Besides interactions between HIV and human proteins, the efficient viral replication requires the interactions between HIV proteins, so-called HIV inter-protein interactions. Although HIV-1 genome only encodes 15 viral proteins, extensive HIV-1 inter-protein interactions have been reported during different stages of the viral life cycle (**Table 1.1**). To our knowledge, there is no review which summarizes HIV-1 inter-protein interactions. To understand the HIV-1 genome-wide interaction, this section summarizes the function of HIV-1 inter-protein interactions reported during the last three decades. Note that our study focuses on HIV-1 as few studies have reported HIV-2 inter-protein interactions (e.g. the p6 positions 15-40 interact with the vpx positions: 73-89 [78]).

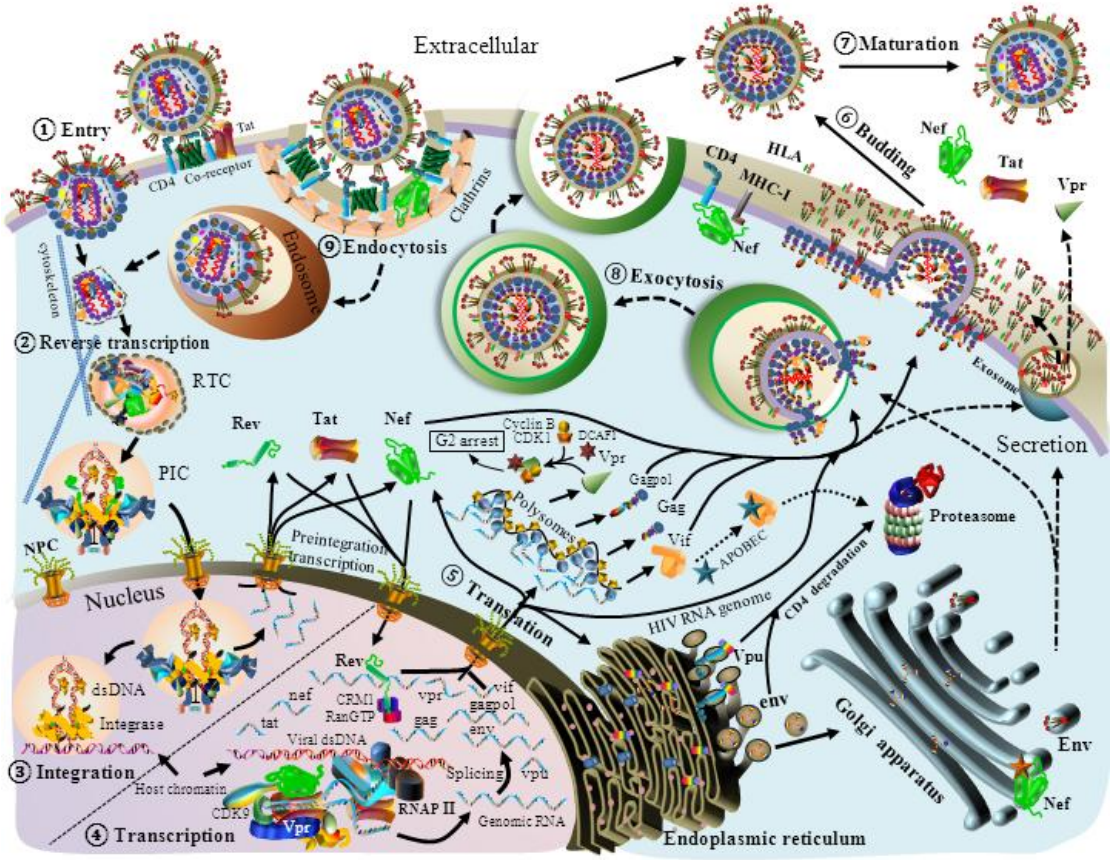


Figure 1.2: Overview of HIV-1 life cycle.

(1) Viral entry (viral fusion): mature virions target host cells through the binding with the cellular receptor CD4 and other chemokine co-receptors (e.g. CCR5, CXCR4).
(2) Reverse transcription: reverse transcriptase in the reverse transcriptase complex (RTC) produces a double-stranded DNA (dsDNA) from a single-stranded RNA [79].
(3) Viral integration: the pre-integration complex (PIC) transports dsDNA into the nucleus through the nuclear pore complex (NPC) [80]. During the viral pre-integration transcription, Rev, Tat and Nef are synthesized from un-integrated dsDNA [81]. Assisted by cellular cofactors (e.g. LEDGF/p75), PIC integrates dsDNA into host chromosome regions with a high transcriptional activity [80].
(4) Viral transcription: viral proteins (Tat, Nef) hijack the cellular transcription machinery to activate the viral mRNA synthesis from the integrated viral DNA [82]. Rev protein exports the viral mRNA and sliced precursors (*vpr*, *vpu*, *nef*, *tat*, *vif*, *env*, *gag*, *gagpol*) into the cytoplasm.
(5) Viral translation: viral mRNAs are translated into precursor proteins in the cellular compartments. Viral mRNAs (*gag*, *gagpol*, *vpr*, *vif*) are translated in cytosolic polysomes [83]. In the cytoplasm, mature Vpr binds with host proteins (DCAF1, CDK1, Cyclin B) to induce the G2 cell cycle arrest. Vif induces the degradation of host APOBEC proteins [84] [85]. Nef plays multiple roles in different cellular compartments through HIV-host interactions [86]. In the rough ER, the cellular proteases (e.g. furin) cleave the *env* glycoproteins into GP120 and GP41, which subsequently assemble into the trimeric complex in the Golgi apparatus via non-covalent interactions [87]. In the Endoplasmic reticulum (ER), the mRNA of Vpu is translated and mature Vpu in the complex with host proteins can retain the newly synthesized CD4 [88]. The dislocated CD4 is subsequently delivered to the

proteasome for degradation [88]. Most *env* proteins are retained in ER and subsequently ubiquitinated and degraded by proteasome [89]. Some of the *env* proteins travel to the Golgi complex to be proteolytically cleaved by the cellular furin or furin-like proteases [90]. Envelope complexes and viral proteins Vpr, Tat, and Nef travel to the extracellular membrane surface via the exosome secretory pathways [91].
(6) Viral assembly and budding: Nascent viral particles are assembled and packaged with two mRNA genomes, viral proteins (Vif, Vpr, *gag*, *gagpol*, *env*) and cellular cofactors (e.g. actin, $tRNA^{Lys3}$, cyclophilin A). Nascent viral particles pinch off from the cellular membrane to infect new host cells [83]. Mature Nef promotes the CD4 degradation to prevent the *Env*-CD4 binding on the extracellular membrane of infected cells [92].
(7) Viral maturation: Protease cleaves Gag and GagPol polyproteins into structural proteins (Matrix, Capsid, Nucleocapsid, p6) and viral enzymes (Protease, RT, Integrase). After the protease-mediated proteolytic process, the immature viral particles become mature for new infections.
(8) Endocytosis: As an alternative pathway, mature virions enter the host cells via cellular endocytosis. Mechanisms that govern HIV endocytosis remain unclear [93].
(9) Exocytosis: As an alternative pathway, nascent HIV virions are released through the exocytosis pathway [94].
Note that protein shapes and sizes are not to scale.

Table 1.1: Summary of HIV-1 inter-protein interaction and their interaction positions

Protein 1,Protein 2	Life stage	Positions in Protein 1	Positions in Protein 2	Ref
GP120, GP41	Entry, Budding, Translation	53,72,73,220,223	542-562	[18]
		36-45, 491- 501		[95]
			593,596,606,610,614,623	[96]
			572,579	[97]
			528,530,552,555, 562,584,608,628	[98]
			593,596,597,601,610	[99]
			501,605	[100]
		225,244		[101]
		66,69,72,73,104,107,109,111, 112,116,213,217		[102]
		44-47,84-86,215-219,222-229, 91-95,241-245,485-492		[103]
			596,597,618	[104]
		382,420,433,438		[105]
		44,53		[106]
		491,494,496,498		[107]
			556,558,563,570,577	[108]
GP41, Matrix	Entry, Budding	Δ 93 C-terminal	12,30	[109]
		201	12,30,34	[110]
		201, Δ 144 C-terminal	49	[111]
		Δ 292-296	34	[112]
		Δ 104 C-terminal	12,30	[113]
		Δ 116-123	18,20,22,32,33	[114]
		Δ 144 C-terminal	8-9, Δ 16-18	[115]

		Δ17 C-terminal		[109]
			62	[116]
			Δ41-57, Δ56-68, Δ67-78	[117]
			13-43	[118]
			Δ116-128	[119]
			18,20,22,29,32,33	[120]
GP120, Tat	Entry	157-171	73-86	[121]
RT, NC	Reverse transcription		1-55	[122]
			12-53	[123]
		Δ13C-terminal in p15	1-71	[124]
RT, Integrase	Reverse transcription; Integration		220-270,243,250,258	[125]
		1-242,387-560	201-288	[126]
			130	[33]
			46-65	[127]
RT, Vif	Reverse transcription		161-164,169-192	[128]
			Δ56C-terminal	[129]
RT, Tat	Reverse transcription		47,49-52	[130]
			Δ60N-terminal	[131]
			1-86	[132]
			49-57	[133]
RT, Nef	Reverse transcription		154-172	[134]
Integrase, Nef			58-206	[135]
Integrase, Rev	Integration		1-30,49-74	[136]
		118-128,66-80	12-23, 53-67	[137]
		66-80,118-128		[138]
Integrase,Matrix	Integration	50-212	132	[139]
Matrix, Vpr	Integration	88-132		[140]
Tat, Vpr	Transcription	50-67	73	[141]
RT, Gag	Budding	183-305		[142]
NC^{Gag}, Vpr	Budding	13-30,34-51	80-96	[143]
NC^{Gag}, Vif	Budding		157-179	[144]
		48-55	68-81,89-100, 162-173,177-189	[145]
$Capsid^{Gag}$, Vif	Budding	219-231		[146]
Nef, P6*	Budding	148-180		[147]
Nef, $GP41^{Env}$	Budding	181-210		[148]
$p6^{Gag}$, Vpr	Budding		84-94	[149]
			1-71	[150]
		32-46		[151]
		35-47		[152]
		15-18		[153]
		34-36		[154]
p6*, Protease	Maturation	65-68		[155]
Protease, Vif	Maturation	1-9		[156]
			41-65	[157]
			78-98	[158]
			81-88,88-98	[159]

1.4.1 Viral entry

In this stage, HIV virions enter into the host cell cytoplasm and initiate the cell infection (**Figure 1.2**). Host cells which express the CD4 receptor on the cell surface

are targeted (e.g. helper T lymphocytes, macrophages, dendritic cells) by trimeric spikes formed by GP120 and GP41 (**Figure 1.3**). Different entry pathways and the involved host factors have been reviewed previously [87, 93, 160-163].

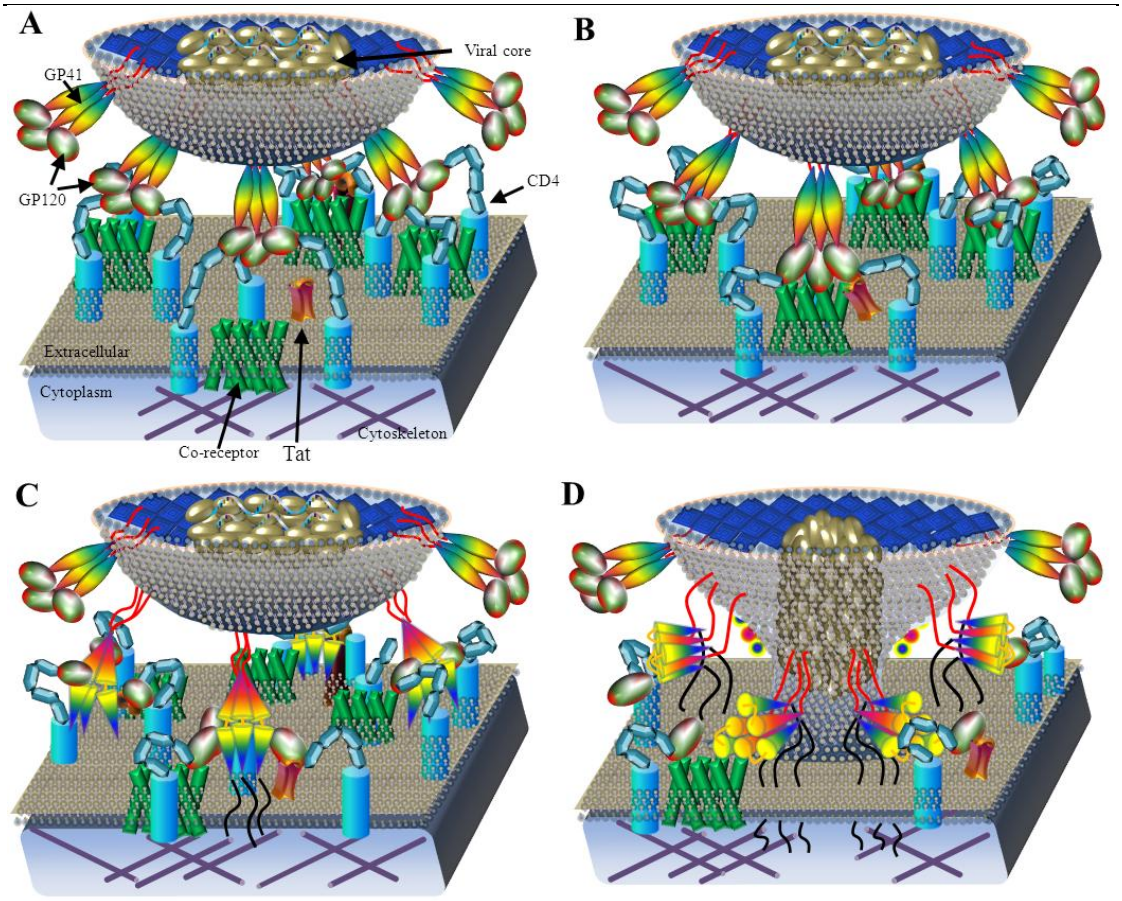


Figure 1.3: Schematic view of HIV-1 inter-protein interaction during viral entry. (A) Viral attachment: GP120 on the mature virion surface binds with CD4 which induces the aggregation of CD4 and co-receptors (e.g. CCR5) [164]. (B) Coreceptor binding: GP120 binds with chemokine co-receptors on the extracellular surface. (C) Interactions between GP120 and co-receptors induce conformation rearrangements in GP120 which expose the GP41 to form six-helix bundles. (D) The six-helix bundles pull viral core to enter the cytoplasm of host cells through the newly-created fusion pore. Protein shapes and sizes are not to scale.

GP120−GP41 protein interaction: GP120 interacts with GP41 to form a trimeric spike through non-covalent interactions, allowing for the flexible structural rearrangement during viral entry [102]. When GP120 binds with the CD4 receptor, the non-covalent interaction between GP120 and GP41 is subsequently disrupted and

a series of conformation changes take place for viral entry [165]. The CD4-induced realignments in viral spikes induce the exposure of GP41 fusion peptide, which is inserted into the cell membrane thereafter (**Figure 1.3**). Early studies of GP120-GP41 interactions showed that mutations in the C5 region of GP120 or the ectodomain of GP41 can disrupt the non-covalent interactions [18, 95, 100, 166]. A Science paper recently reported the first atomic-level structure of GP120-GP41 trimers, showing that an invariant 7-stranded β-sandwich in GP120 maintains GP120-GP41 interactions and regulates GP41 transitions [18].

GP41^{Env}−Matrix^{Gag} protein interaction: Mutagenesis, biochemical and biophysical assays have shown that the N and C terminus of HIV-1 Matrix in Gag precursors (Matrix^{Gag}) interact directly with the cytoplasmic tail of GP41 (GP41CT) [116]. GP41CT plays an essential role for the Env incorporation during viral budding [111, 112, 167] and for the pre-bundle rearrangement of Env protein during viral entry [114, 168, 169]. Major functions of GP41-Matrix interaction include: (1) Env proteins are incorporated into nascent virions through the interaction between GP41^{Env} and Matrix^{Gag} [108, 110, 111, 113, 116, 167]. (2) The maturation of viral core is associated with the activity of GP41CT [170]. (3) The stabilization of Env glycoproteins is enhanced by the GP41^{Env} -Matrix^{Gag} interaction in a cell-dependent manner [120, 169, 171].

GP120−Tat protein interaction: Tat can bind with GP120 to enhance viral entry [121, 172]. This interaction involves several processes. Firstly, Tat is released to the extracellular space by infected cells [121, 173, 174]. Secondly, on the extracellular membrane of infected and uninfected cells in neighboring areas, the released Tat binds to chemokine receptors CCR2 and CCR3 but not CCR1, CCR4, and CCR5 [175]. Thirdly, extracellular Tat directly interacts with GP120 on the cell surface, while Tat is dispensable for viral entry [121]. Moreover, the GP120-Tat interaction impacts on viral entry but not on Tat-mediated transactivation. When considering interaction domains, residue positions in the V1/V2 loop of GP120 interact with the second exon of Tat [121]. It remains unclear how GP120-Tat interaction exerts an influence on the conformation rearrangement of GP120.

No interaction between GP41 and Vif: An early study suggested that Vif may interact with GP41CT [176]. Two studies later showed that the GP41CT activity was independent of Vif, ruling out the possible GP41-Vif interaction during viral entry [177, 178].

1.4.2 Reverse transcription

During reverse transcription, reverse transcriptase produces a double-stranded DNA from the single-stranded viral RNA [179-181]. After the viral entry, a series of processes take place to form the reverse transcriptase complex (RTC) in viral core (**Figure 1.4**). Where and when the reverse transcription happens is still debatable but recent evidence favors the hypothesis that reverse transcription takes place in the intact capsid core and is triggered by massive deoxyribonucleotides in the cytoplasm after viral entry [181-183]. During reverse transcription, the intact capsid core moves toward the nuclear pore on microtubules and RTC is turned into the pre-integration complex (PIC) [184] (**Figure 1.2**). Different aspects of HIV reverse transcription have been reviewed such as the enzymatic function of reverse transcriptase [34, 182], the maturation of RTC [79], the strand transfer process and recombinant events [180]. Here we focus on HIV-1 inter-protein interactions.

RT−Integrase protein interaction: Integrase can physically interact with RT to enhance the initiation of reverse transcription [125, 185-188]. Mutagenesis analysis showed that Integrase mutants could severely impair reverse transcription and weaken the RT-Integrase interaction [33, 186]. Regarding RT-integrase interaction domains, the C terminus of Integrase (positions: 220-270) is important to bind with RT [33, 186, 188]. The interaction domain in RT covers a board area including the finger-palm domain (positions: 1-242) and the C terminus of the connection subdomain (positions: 387-560) [33]. It is now agreed upon that Integrase can interact with RT but the role of this interaction remains unclear. It has been shown that Integrase stimulates both initiation and elongation at the early steps of reverse transcription, while integrase exerts no effect on steps at or before template-primer annealing [189]. Integrase exerts no influence on the RT processivity, whereas RT can stimulate the Integrase-mediated strand transfer reaction in a concentration-dependent manner [126]. In another study, RT however inhibited the Integrase strand transfer reaction in

the cell-free assays [186, 188]. This discrepancy might be due to different HIV-1 strains, cell lines or experiments. Further studies are still needed to address the functional interactions between Integrase and RT.

RT−Nucleocapsid protein interaction: At the initiation stage of reverse transcription, RT interacts with Nucleocapsid to anneal the $tRNA^{Lys3}$ onto the primer binding site of HIV genomic RNA [123, 190]. During reverse transcription, the RT-Nucleocapsid interaction can: (1) promote the RT activity by reducing the rate of incorrect nucleotide incorporation [122, 191, 192]; (2) improve the ribonuclease activity of RNase H domain in RT [185]; (3) counteract the decreased strand transfer efficiency caused by RT mutations [124]. Two zinc fingers in Nucleocapsid are important to interact with the C terminus of the RNase H domain in RT [123, 124, 190, 192].

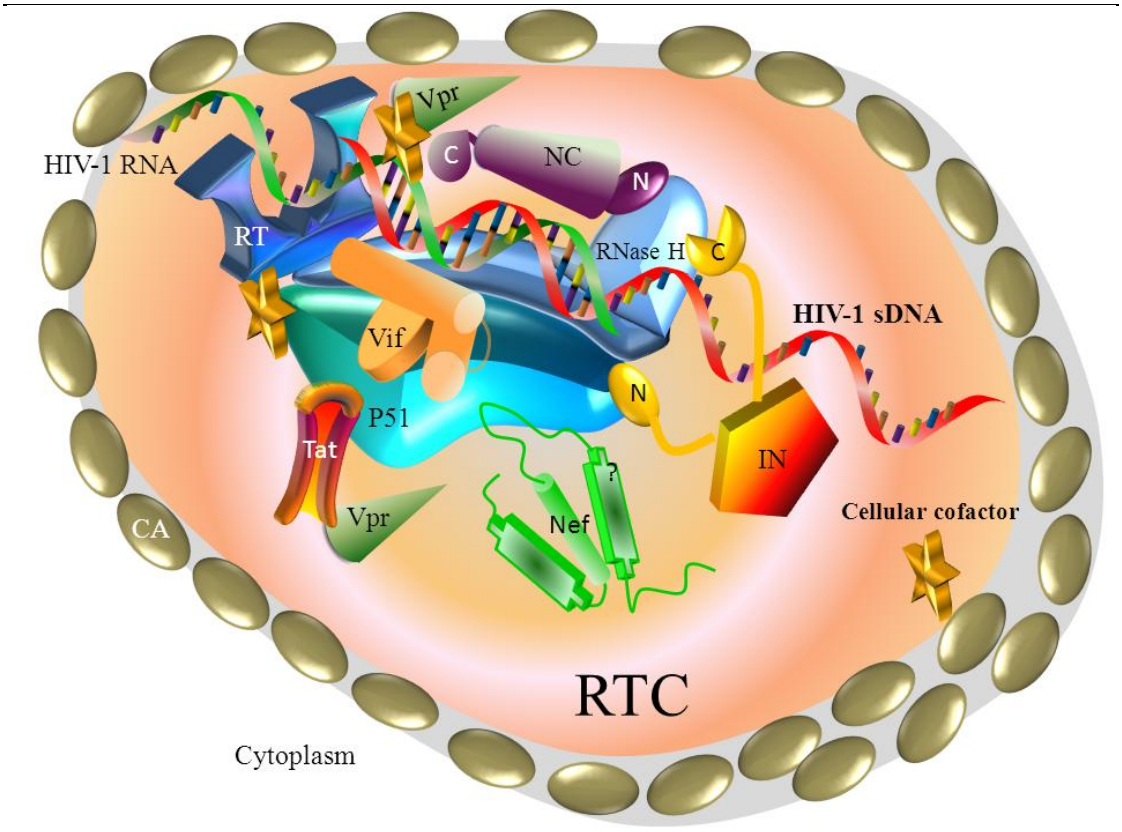


Figure 1.4: Schematic view of HIV-1 inter-protein interactions in reverse transcriptase complex. During reverse transcription, RTC can produce viral dsDNA, protect dsDNA from nuclease digestion and prevent viral self-integration [183]. Cellular cofactors can interact with RTC to facilitate reverse transcription [74].

Question marks indicate unclear interaction positions. Note that protein shapes and sizes are not to scale.

RT−Nucleocapsid protein interaction: At the initiation stage of reverse transcription, RT interacts with Nucleocapsid to anneal the $tRNA^{Lys3}$ onto the primer binding site of HIV genomic RNA [123, 190]. During reverse transcription, the RT-Nucleocapsid interaction can: (1) promote the RT activity by reducing the rate of incorrect nucleotide incorporation [122, 191, 192]; (2) improve the ribonuclease activity of the RNase H domain in RT [185]; (3) counteract the decreased strand transfer efficiency caused by RT mutations [124]. Two zinc fingers in Nucleocapsid are important to interact with the C terminus of RNase H domain in RT [123, 124, 190, 192].

RT−Vif protein interaction: RT was found to interact with Vif using the glutathione-S-transferse (GST) pulldown assay [128]. Vif is a component of RTC [129] and PIC [193]. A reduced RT activity was observed in the presence of Vif mutant [194], and the dsDNA synthesis is harmed by Vif defects in non-permissive cells [195, 196]. During the early stage of reverse transcription, the RT-Vif interaction can stimulate the $tRNA^{Lys3}$ primer annealing by increasing the polymerization rate, decreasing the pausing of reverse transcription during ssDNA synthesis and increasing the RT processivity [129]. The C terminus of Vif can interact with RT [128], but the interaction domain in RT remains unclear.

RT−Tat protein interaction: Direct RT-Tat interaction was confirmed by GST pull-down and immune-precipitation assays [133]. A ratio of at least 2:1 between Tat and RT is required for the increased RT activity [197]. While the RT-Tat interaction is crucial for reverse transcription, biological mechanisms and interaction domains remain unclear. Tat can stimulate reverse transcription [130] and virus lacking Tat cannot initiate reverse transcription efficiently [131], whereas other studies have reported that Tat can suppress reverse transcription in cell-free assays [131, 132, 198]. Further studies are still needed to address the role of RT-Tat interaction and corresponding interaction domains.

RT−Nef protein protein interaction: Direct interaction between RT and Nef was confirmed by co-precipitation assays [134]. The average ratio of Nef to RT is

estimated to be 1:10 in HIV-1 particles [199]. The RT-Nef interaction enhances the binding affinity of RT to RNA, a process which is independent of Nef binding to RNA [134]. Based on mutagenetic analyses, the p51 in RT was found to interact with the disorder loops in the C terminus of Nef [134].

RT−Vpr protein interaction: Peptides derived from Vpr (positions: 57-71 and 61-75) can physically interact with RT and inhibit reverse transcription [200]. While direct evidence of full-length Vpr and RT interaction is still lacking, previous studies have shown that: (1) Vpr interacts with the $tRNA^{Lys3}$ synthetase to influence the initiation of reverse transcription [201]. (2) Vpr is associated with the packed filaments in RTC [202]. (3) Vpr is cosedimented with the synthesized viral dsDNA [203]. (4) Both Vpr and RT are packed into the viral core. Further studies are needed to address the functional interactions between RT and Vpr.

1.4.3 Viral integration

As illustrated in **Figure 1.5**, HIV-1 integration involves: (1) the 3'-end processing, Integrase removes two nucleotides at the 3' end of dsDNA in cytoplasm. (2) Nuclear import, PIC containing dsDNA is imported from cytoplasm into nucleus through the nucleus pore complex. (3) PIC targets the host chromosome domains with a high transcriptional activity. This process is assisted by cellular cofactors such as LEDGF/p75 – a cellular transcriptional coactivator serving as a tethering protein between the PIC and host chromosomes. (5) Strand transfer reaction: viral dsDNA is inserted into the host chromosomes via the strand transfer reaction exerted by viral Integrase. (6) Gap repair, the unpaired regions of DNA between HIV dsDNA and host dsDNA are repaired by cellular cofactors. Further details about nuclear import, pre-integration transcription and host proteins (e.g. LEDGF/p75, INI1) have been reviewed in [81, 204, 205].

Integrase−Matrix protein interaction: The C terminus of Tyrosine-phosphorylated Matrix was reported to interact with the central domain of Integrase using immune-precipitation and western-blot assays [139]. The Integrase-Matrix interaction facilitates the nuclear localization of viral dsDNA in the absence of mitosis [139]. Tyrosine phosphorylation of Matrix was necessary to interact with Integrase [139]. Replacing tyrosine with phenylalanine at the residue position 132 of Matrix can block

the nucleus import [139]. The core domain of Integrase (positions: 50-212) is important for the Matrix binding [139].

Integrase−RT protein interaction: Integrase can physically interact with RT [33, 125, 127, 187]. Two important roles of the Integrase-RT interaction have been reported during viral integration. First, RT serves as a regulator to inhibit the disintegration activity of Integrase [206]. Second, RT in PIC may inhibit the strand transfer activity of Integrase to prevent auto-integration before viral dsDNA reaches the host chromosome [187]. Auto-integration is a suicidal process that viral dsDNA integrates within itself [207]. Considering the interaction domains, the C terminus of Integrase is necessary and sufficient to interact with RT [33, 125]. Mutations (W243E, V250E, K258A) in the C terminus of Integrase could severely impair the Integrase-RT interaction [125]. Moreover, RT positions (L168, F171, Q174, I178) may interact with Integrase [127]. Notably, different interaction domains have been observed during viral integration and reverse transcription (**Table 1.1**). Further studies are needed to verify whether this difference is due to different cell-lines, HIV-1 strains or conformation switch.

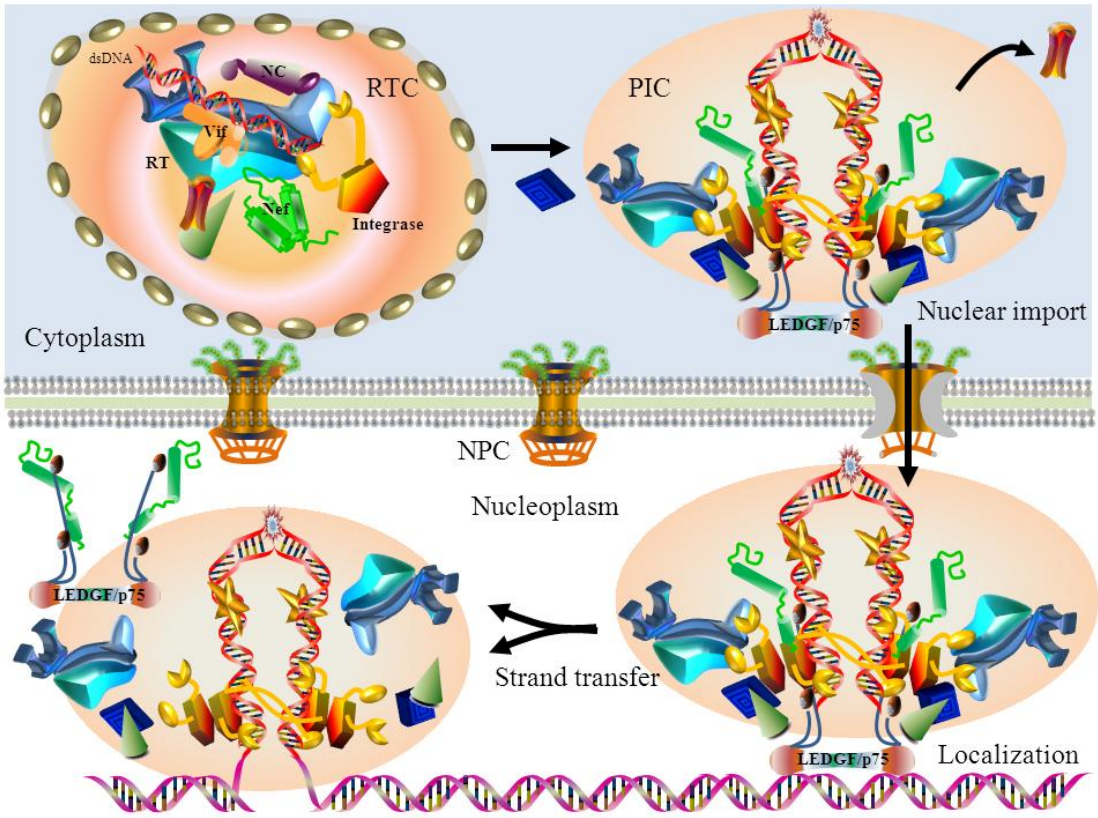


Figure 1.5: Schematic view of HIV-1 inter-protein interactions during viral integration. (A) Viral uncoating: RTC turns into PIC with the recruitment of Matrix,

Rev and host factors (e.g. LEDGF/p75). PIC contains host factors and viral Integrase, Matrix, RT and Vpr and little to no Capsid. Meanwhile, Nef, Nucleocapsid, Tat and most Capsid proteins are dissociated from PIC [204]. (B) Nuclear import: PIC is imported from cytoplasm to nucleoplasm through the nuclear pore complex. (C) Chromosome localization: PIC is tethered to the targeting chromosome via the assistance of LEDGF/p75. (D) Integration: PIC integrates viral dsDNA into host chromosome. Question marks indicate unclear interaction positions. Note that protein shapes and sizes are not to scale.

Matrix−Vpr protein interaction: Matrix was found to directly interact with Vpr in HIV-1 mature virions using coimmunoprecipitation and two-hybrid GAL4 assays [140]. The Matrix-Vpr interaction may improve the stoichiometry of nucleophilic components in PIC and promote the nuclear import of dsDNA in non-dividing cells (e.g. monocyte-derived macrophages) [208]. The C terminus of Matrix may interact with Vpr [208], while the interaction domain in Vpr remains unclear.

1.4.4 Viral transcription and translation

After HIV-1 dsDNA is integrated into host chromosomes, viral regulatory proteins Tat, Rev and Nef can interact with cellular miRNA machinery to control the gene expression for the virus production (**Figure 1.6**). Previous studies have reviewed the biological processes of HIV-1 transcription in different cell lines [209] and the interaction between HIV proteins and cellular transcriptional factors [82, 210].

Tat−Vpr−cyclin T1/CDK9 complex: Vpr can interact with both Tat and cyclin T1 [141]. Tat is a regulatory protein known to interact with cyclin T1 and its partner, CDK9, to promote viral transcription. In the presence of Vpr, a super-activation of the long terminal region (LTR) by Tat and cyclin T1/CDK9 was observed [141]. Vpr can strongly activate theHIV-1 LTR when both cyclin T1/CDK9 and Tat are present [141]. Vpr is known to the pre-integration transcription which synthesizes Rev, Tat and Nef before viral integration (**Figure 1.2**). Tat is dispensable for the Vpr-mediated activation of pre-integration transcription [211]. Moreover, the residue position R73 in Vpr may interact with Tat [141], while the interaction domain in Tat remains unclear.

Tat−Nef protein interaction: Physical interaction between Tat and Nef was identified using transient transfection, co-immunoprecipitation and GST pull-down

assays [212]. Both Tat and Nef are expressed before viral integration and localized in nucleus (**Figure 1.2**). Nef can enhance the Tat-mediated transactivation of HIV-1 LTR [212]. On the cell membrane, Nef promotes the Tat-mediated viral transcription via an hnRNP-K-Nucleated signaling complex [213-215]. Nef exerts an influence on the Tat-mediated gene expression either by the direct interaction or the signaling pathways mediated by cellular cofactors. Interaction domains between Tat and Nef remain unclear.

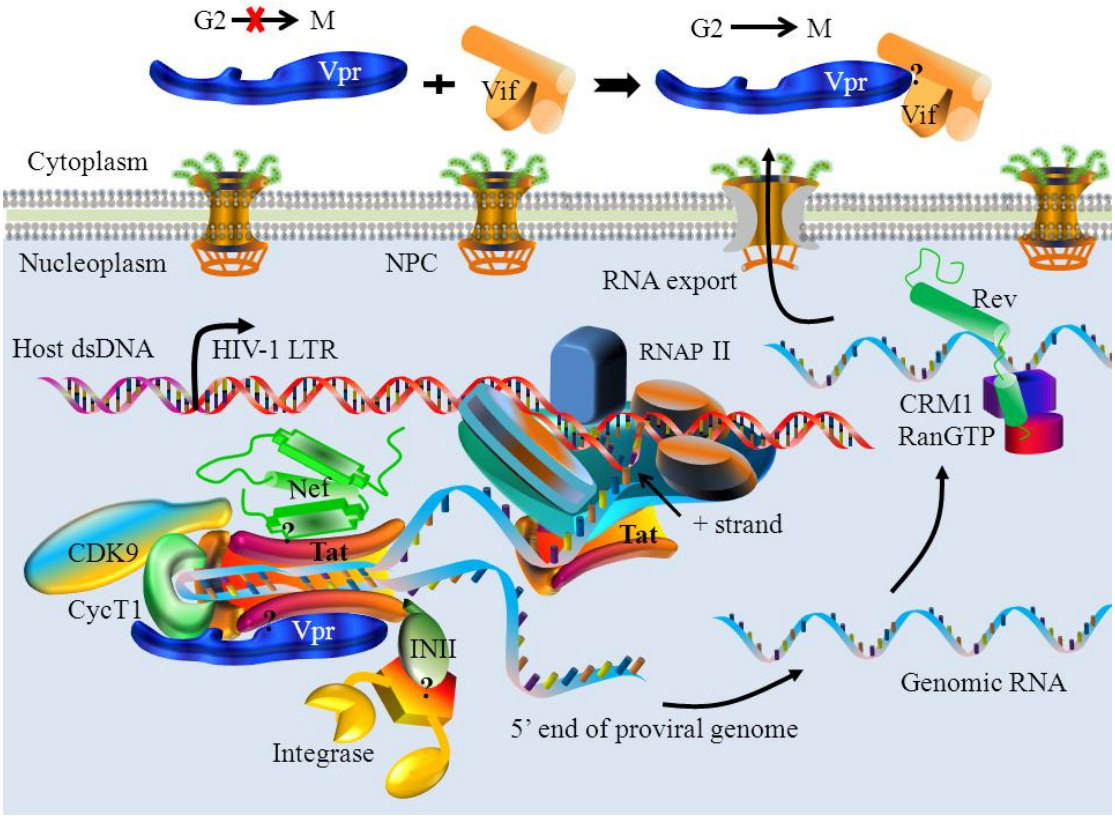


Figure 1.6: Schematic view of HIV-1 inter-protein interactions during viral transcription. To initiate viral transcription, the transactivator Tat binds to the TAR of RNA which is a regulatory element located at the downstream of HIV-1 LTR. Tat can recruit positive transcription elongation factors (e.g. Cyclin T1 (CycT1), cyclin-dependent kinase 9 (CDK9)) to form a transcription complex. This complex then activates the CDK9 kinase leading to the hyperphosphorylation of RNA polymerase II (RNAP II). The hyperphosphorylated RNAPII interacts with Tat and other transcription elongation factors to transcribe the viral genomic RNA [216]. After the transcription of genomic RNA, Rev cooperates with the cellular factors Crom1 and RanGTP to export the genomic and sliced RNAs into cytoplasm. Question marks indicate unclear interaction positions. Note that protein shapes and sizes are not to scale.

Vpr−Vif protein interaction: Vif can physically interact with Vpr to degrade Vpr via the ubiquitin/proteasome pathway [217]. The downgradation of Vpr reduces the accumulation of infected cells at the G2 cell cycle arrest [217]. The interaction domains between Vpr and Vif remain unclear.

Tat−INII−Integrase association: Direct interaction between Tat and Integrase has not been reported. Tat may associate with Integrase through a chromatin remodeling factor integrase interactor I (INII), also known as hSNF5. Being recruited into PIC [218], INII interacts with Integrase to enhance viral integration [219] and interacts with Tat to enhance the Tat-mediated transcription [220].

1.4.5 Viral budding

As illustrated in **Figure 1.7**, the viral genomic RNA, envelope proteins and accessory proteins are assembled into nascent virions, which subsequently pinch off from extracellular membrane [221]. Previous studies have reviewed *env* trafficking and incorporation [222], HIV-1 genome packaging [223], APOBEC pathway [224], membrane lipids [225] and cellular cofactors in promoting viral budding [226]. Here we focus on the HIV-1 inter-protein interactions during viral budding.

Vif−NC^{Gag} protein interaction: Vif can interact with nucleocapsid coded in the Gag precursor (NC^{Gag}) [144, 145, 227, 228]. This interaction has been reported to impact the HIV-1 life cycle in three aspects: (1) Vif can inhibit the hybridization of tRNA^{Lys3} and the Nucleotide-mediated formation of RNA dimers [229]; (2) Vif inhibits the protease-mediated proteolytic processing at the cleavage site between p2 and Nucleocapsid [230]. (3) NC^{Gag} becomes less stable in the viral core when Vif is absent [231]. The motifs at the C terminus of Vif interact with Gag [144], while the interaction domain in Nucleocapsid remains unclear.

Vpr−p6^{Gag} protein interaction: Vpr is incorporated into nascent HIV particles through its interaction with the p6 domain of Gag precursors (p6^{Gag}) [149, 232, 233]. The Vpr-p6^{Gag} interaction allows for the Vpr incorporation on the lipid bilayer membrane [150, 234]. The motif near the C terminus of p6 [151-154] can interact with Vpr (position: 1-71 [150], 84-94 [231]).

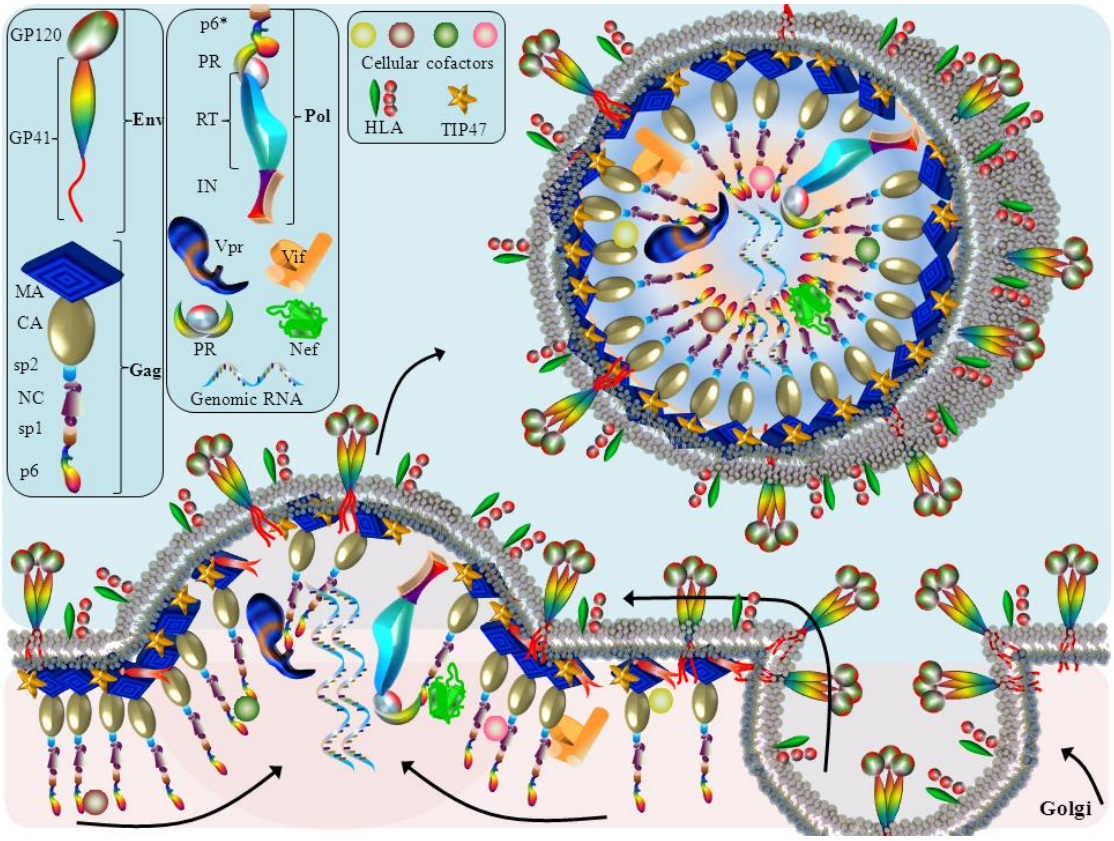


Figure 1.7: Schematic view of HIV-1 inter-protein interactions during viral budding. Env trimmers are exported to extracellular membrane through a secretory pathway. Gag targets to the membrane rafts where sphingolipids, glycolipids and cholesterol are rich. (1) Matrix^{Gag}-GP41 interaction: the N-myristoylation domain of Matrix^{Gag} interacts with GP41, allowing the incorporation of Gag into HIV particles. This interaction is mediated by the cellular cofactor TIP47 [235]. (2) Gag-Vif interaction. NC^{Gag} and Matrix^{Gag} can interact with Vif to assist the incorporation of Vif into HIV particles. (3) Vpr-p6^{Gag} interaction: Incorporation of Vpr into nascent virions depends on the Vpr-p6^{Gag} interaction. (4) NC^{Gag}-Vpr interaction: NC^{Gag} can cooperate with p6^{Gag} to incorporate Vpr into HIV particles. This interaction also promotes the Vpr-RNA interaction during the RNA encapsulation. (5) Nef-p6*^{Gagpol} interaction assists the Nef incorporation. Note the protein shapes and sizes are not to scale.

Vpr−NC^{Gag} protein interaction: Vpr interacts with Nucleocapsid in the Gag precursor (NC^{Gag}), which allows for the incorporation of Vpr into viral particles [143, 236]. NC^{Gag} promotes the Vpr-RNA interaction during the encapsulation of genomic RNA [237]. Mutagenesis analysis has shown that the zinc fingers of NC^{Gag} [143] interact with the C terminus of Vpr (positions: 70-80) [237].

Env−Nef protein interaction: Nef can physically interact with the GP41 cytoplasmic tail (GP41CT) [148]. Deletion of GP41CT abrogates the Nef-induced

increase of viral infectivity in CD4+ lymphocytes [148]. Mutagenesis analyses have shown that the C terminus of Nef (positions: 181-210) interacts with GP41CT [148]. Nef may improve HIV-1 infectivity by enhancing the Env incorporation because Nef induces the downregulation of CD4 to prevent Env-CD4 binding [92]. Moreover, Env protein is dispensable for the Nef-induced viral infectivity [238], whilst the accessory protein Nef is dispensable for viral budding and entry [239].

Nef−p6* protein interaction: p6* is a transframe peptide region which separates the Gag nucleocapsid domain from Protease. Nef interacts with the p6* domain of GagPol precursors in the intermediate compartment between the ER and trans-Golgi networks [147]. This interaction allows for Nef incorporation into nascent HIV particles [147, 240]. Moreover, the flexible loop in Nef (positions: 148-180) may interact with p6* [147]. Nef mutants without this flexible loop cannot interfere with the processing of GagPol polyprotein and do not incorporate into HIV particles [147].

Matrix−GP41 protein interaction: Mediated by the cellular cofactor TIP47, the GP41 cytoplasmic tail (GP41CT) can interact with the N terminus of Matrix for Env incorporating into HIV particles [114, 235, 241]. Gag and Env proteins are colocalized in the plasma membrane and Golgi apparatus [242]. Matrix-GP41CT interaction is critical for the envelope association with lipid rafts in the extracellular membrane [243]. The myristoylation of the N terminus in Matrix is required for Env incorporation [244].

Vpu−Matrix^{Gag} protein interaction: Vpu interacts with Matrix in the Gag precursor (Matrix^{Gag}) [245]. This interaction may enhance the binding of Gag to the extracellular membrane [245]. The N terminus of Matrix^{Gag} is required to interact with Vpu [245], while interaction domain in Vpu remains unclear.

RT−Gag protein interaction: RT is incorporated into virus-like particles through the RT-Gag interaction [142]. The thumb domain of p51 is required for RT incorporation, whilst the matrix and p6 in Gag precursors interact with RT [142]. Yet, it is rare for the production of mature RT in cytoplasm, as well as the transport of mature RT to the extracellular membrane.

1.4.6 Viral maturation

After viral budding, immature virions undergo the maturation process during which multimerized Gag precursors are cleaved into mature proteins using viral protease [246]. Viral enzymes (protease, RT, integrase) in the GagPol precursors are not functional and must be cleaved and folded for viral activity [247]. Further details about the maturation process, the core morphology and structure can be found in [142, 246]. Here we focus on HIV-1 inter-protein interactions observed during viral maturation.

Protease−Vif protein interaction: Protease can directly interact with Vif [156]. This interaction can: (1) interfere the Protease dimerization [248]; (2) inhibit the Protease-mediated proteolytic processing at the p2-Nucleocapsid cleavage site [230]; (3) prevent the Protease-mediated digestion of cellular proteins [248]. Moreover, the N terminus of Protease (positions: 1-9) interacts with the central domain of Vif (positions: 78-98) [248]. Mutations at Vif positions (36,47,101,117,124) were associated with protease drug resistance [249]. Vif-derived peptides (positions: 81-88, 88-98) can inhibit the activity of Protease [159].

Protease−RT protein interaction: A direct interaction between Protease and RT was identified using immunoprecipitation, Western blot experiments and an enzyme-linked immunosorbant assays [250]. Previous studies have shown that: (1) RT increases the catalytic activity of Protease without interfering with the protease dimerization [251]. (2) RT increases Protease activity in a dose-dependent, pH-dependent and concentration-dependent manner [252]. (3) Protease inhibits the catalytic activity of RT whereas the activity of RNase H is not affected in the presence of Protease [250]. The interaction domains remain unclear.

Integrase−Nef and Protease−Nef interactions: A physical interaction between Integrase and Nef has consistently been detected by different assays [135]. RT was reported to interact with Nef using GST pull-down and yeast two-hybrid assays. Protease-Nef was reported using yeast two-hybrid and immunoprecipitation assays. However, the biological function of Integrase-Nef interaction is unknown, neither interaction domains.

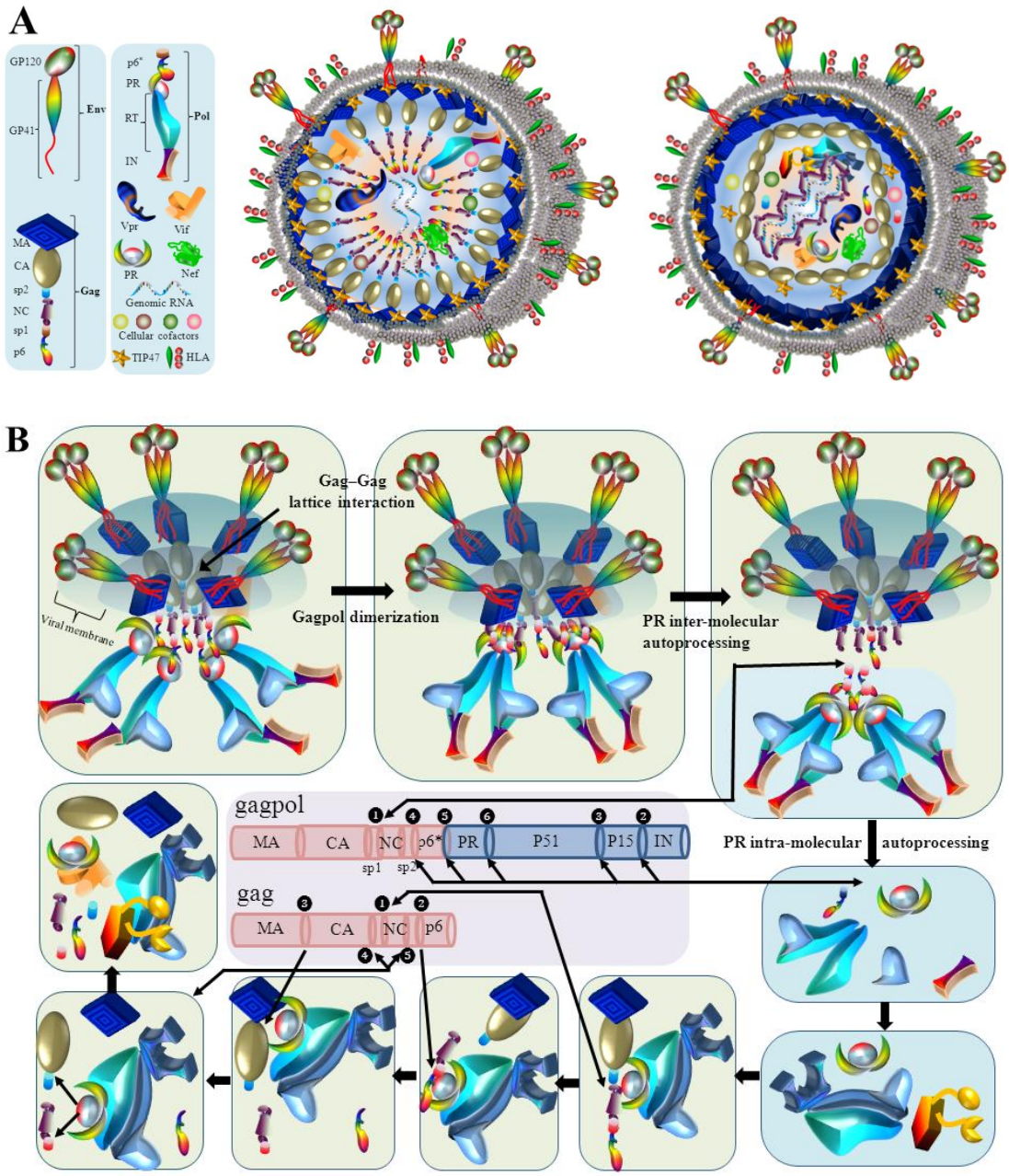


Figure 1.8: Schematic view of HIV-1 inter-protein interactions during viral maturation. (A) Schematic view of immature (left) and mature (right) HIV-1 particles. Immature particles become mature after protease-mediated autoproteolytic processing. (B) Viral maturation. Seven steps are observed: (1) GagPol precursors aggregate in proximity using the Gag-Gag lattice interaction. GagPol dimerization within two Gagpol precursors induces the construction of Protease dimers with a low enzymatic activity [20]. (3) Protease intra-molecular autoprocessing: the cleave site between Nucleocapsid and sp1 is cleaved by Protease within the same Gagpol dimer. (4) Protease inter-molecular autoprocessing: Protease in one Gagpol dimer cleaves the sites (p6*-Protease, PR-p51, p51-p15, p15-IN) in the other Gag-Pol dimer. This process results in the production of Protease dimers and monomer HIV proteins (Integrase, p51, p15, p66). (5) Maturation of HIV-1 enzymes. Integrase and RT are folded and become mature. (6) Protease-mediated proteolytic processing. Mature Protease cleaves Gag precursors in a specific order: sp2-NC → sp1-p6 → MA-CA →

CA-sp2 → NC-sp1. This process frees MA, CA, sp1, sp2, NC and p6 from Gag precursors, during which RT interacts with Protease to enhance the proteolytic processing. The incorporated Vif binds with Protease dimers to inhibit the digestion of cellular cofactors. (7) After the maturation of HIV-1 structural proteins, a series of conformation changes turn immature particles into mature particles. Note that 50 to 63 native HLA-II complexes are in the membrane of HIV-1 particles [253]. One HIV particle has an average of only 7–14 spikes [254], ~1400 Gag polyproteins [255], 30−80 Vif proteins [256]. The protein shapes and sizes are not to scale.

Integrase−Vpr association: While both Vpr and Integrase are components of PIC, a physical interaction between Integrase and Vpr has not been reported. Evidence has shown that: (1) Vpr can stimulate the strand transfer reaction of Integrase [257]. (2) Vpr directly binds to DNA or RNA in a non-specific manner [257] and enhances the binding of Integrase to viral DNA in PIC [237]. (3) The full-length Vpr or the C terminus of Vpr (positions: 52-96) can inhibit viral integration [257]. (4) Vpr-derived peptides (positions: 57-71, 61-75) can inhibit the activity of Integrase via a direct interaction [200]. Further investigations are still needed to show whether Integrase interacts with Vpr.

Capsid−Integrase association: Direct interaction between Capsid and Integrase has not been reported. Yet, Integrase is required to sustain the interaction between Capsid and cyclophilin A, a cellular peptidyl-prolyl isomerase [258]. The stability of viral core formed by capsid was decreased in the presence of Integrase mutants (e.g. C130S). Mutations in Capsid exert a deleterious impact on both nuclear targeting and integration [259]. Nevertheless, a direct interaction between Capsid and Integrase remains unclear.

1.4.7 Summary

Investigation of HIV inter-protein interactions can enrich our understanding on HIV genome-wide interaction and evolution, possibly providing useful information on HIV vaccine and drug design. This section summarizes the known HIV-1 inter-protein interactions reported in the last three decades. Given that only 15 proteins are encoded in the HIV genome, a high level of HIV inter-protein interactions is expected. Indeed, 15 HIV-1 proteins have shown different capacities to interact with each other during the viral life cycle (**Figure 1.9**). Accessory proteins (e.g. Nef) usually take multiple

functions with more opportunities to interact with others, while structural proteins (e.g. Capsid) are restricted to structural functions with lesser interactions. HIV inter-protein interactions in different stages of HIV life cycle provide many possible mechanisms to enhance viral replication. For instance, the interactions between Protease and Gag (Chapter 3), between GP41 and Matrix [260] can cause protease drug resistance, suggesting the crucial role of HIV genome-wide interaction and evolution. Nevertheless, detecting transient states of protein interactions has been proven difficult and the ideal experiments designed to capture these events will likely be complex and expensive.

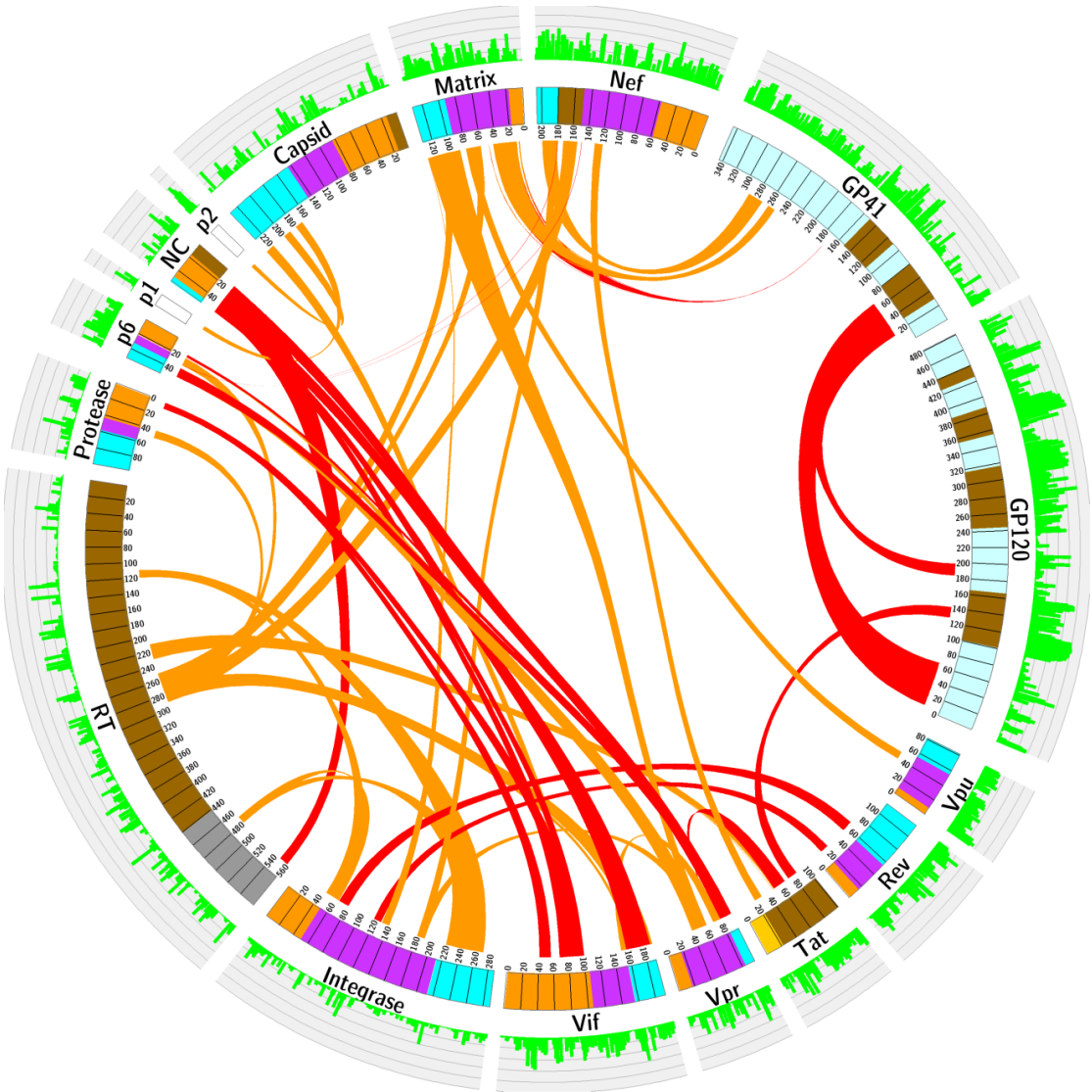


Figure 1.9: Mapping of the documented inter-protein interactions at the HIV-1 genome. In the center, red links indicate the interaction domains between two proteins. Orange links indicate the direct interactions between two HIV-1 proteins but interaction regions have not been resolved (**Table 1.1**). In the circle, the N terminus, the central domain and the C terminus of HIV-1 proteins (Matrix, Capsid, NC, p6,

Protease, Integrase, Vpu, Rev, Vif and Vpr) are colored orange, purple and sky blue, respectively. The p51 and p15 domains in RT are colored brown and gray, respectively. In the GP120 protein region, the four variable domains (V1/V2, V3-V5) and five conserved domains (C1-C5) are colored by brown and sky blue, respectively. In GP41, two helix regions are colored brown and the others colored sky blue. On the out layer, the amino acid genetic diversity of HIV-1 subtype B genome is shown in green. The diversity values between 0 and 1 are mapped on five sub-layers. The data of protein interaction domains are available in **Table 1.1**. The genetic diversity data of HIV-1 subtype B genome is described in Chapter 2.

Investigation of HIV genome-wide interactions faces difficulties in designing experiments that accurately detect the interaction complexes of HIV proteins during the different stages of the viral life cycle. HIV protein can use their intrinsically disordered structures to interact with other HIV proteins in transient states [29]. Because of the dynamic protein structures in different biological contexts, biochemistry experiments performed in cell-free settings may have underestimated the nature of potential interactions. For instance, before the viral maturation, HIV-1 protein precursors (e.g. Gag) have flexible structures with varied interaction properties. After viral maturation, the products of Gag have relatively conserved domains to construct stable viral particles (**Figure 1.8**).

Peptide inhibitors have been proposed to mimic the interaction domains by prohibiting protein interactions, leading to the inhibition of viral replication (Chapter 2). Identification of key interaction domains in HIV-1 proteins may lead to new inhibitors with novel mechanisms of action. Although many HIV-1 inter-protein interactions have been identified in different experimental assays, the majority of the possible interaction domains still remain unclear (**Table 1.1**). We indeed observed discordance results of the biological mechanisms and interaction positions reported in different studies. Further studies are still needed to clarify the unclear protein interactions. Moreover, it has been shown that HIV-2 proteins may play different roles compared to HIV-1, while very few studies have investigated the HIV-2 protein interactions. It remains a challenge to explore the genome-wide interactions in both HIV-1 and HIV-2.

While detecting HIV genome-wide interactions using the biochemistry experiments are expense and time-consuming, bioinformatics methods may provide alternative strategies. For instance, many protein-protein interaction models have been proposed in the last two decades. Sequence-based statistical analyses have also shown promising performance (Chapter 4, 5). Armed with large-scale sequence datasets and literature results, it is possible to model the genome-wide networks and explore all

potential interactions and coevolution at the HIV-1 genome. In this thesis, we conducted interesting projects and contributed to this research area.

1.5 HIV antiretroviral treatment

To date, a curative drug or effective vaccine to treat HIV is still uncertain. A high diversity of HIV strains exerts a difficulty in developing a global HIV vaccine [261]. During the past three decades, many attempts have been made to develop a vaccine with a protective efficacy. Novel vaccine strategies and immunologic principles have been proposed [261-263]. As of May 2014, only three candidate vaccines have completed phase-III clinical trials (http://www.iavireport.org/). At the cost of 119 million dollars, the RV144 vaccine trial conducted in Thailand was the only one trial which demonstrated that an HIV-1 vaccine may elicit a modest and transient protection against the HIV-1 acquisition [261]. To cover a wide range of challenging strains, RV144 tested the "prime-boost" combination of two vaccines: ALVAC® HIV vaccine (the prime) and AIDSVAX® B/E vaccine (the boost) [264]. ALVAC is comprised of a canarypox virus vector engineered with three HIV-1 genes (*gag*, *pol*, *env*). Canarypox is a bird virus which cannot cause disease or replicate in humans [265]. In October 2003, a total of 16402 HIV negative volunteers aged 18-30 participated in the randomized double-blind study groups receiving the prime-boost combination. A follow-up HIV test was conducted in July 2006. The results showed 51 infections in the vaccine group compared to 74 in the placebo group, which indicates a 31.2% (95% confidence interval: 1.1 to 52.1%, p-value=0.04) reduction of HIV-1 acquisition. As debated intensively in the HIV community, RV144 demonstrated a very slight but statistically significant degree of vaccine efficacy [266]. Moreover, a follow-up study showed no difference between the vaccine and placebo groups after 2.5 years [266]. Overall, the design of HIV vaccine still poses as one of the most difficult challenges in the coming years.

With limited achievement from vaccine studies, extensive HIV research has been oriented to the discovery of anti-HIV drugs during the last three decades. Until May 2014, over 25 drugs have been approved by FDA to treat HIV infected patients in clinical practice [267]. Most approved inhibitors are small molecules, except for one peptide inhibitor (T20). Most anti-HIV inhibitors have been designed to target viral

enzymes and envelope proteins (**Figure 1.2**). As shown in **Table 1.2**, the FDA-approved inhibitors are classified into 5 different drug classes (see reviews [267, 268]): (1) nucleoside analog reverse-transcriptase inhibitor (NRTI), (2) non-nucleoside analog reverse-transcriptase inhibitor (NNRTI), (3) protease inhibitor (PI), (4) integrase inhibitor, and (5) entry (fusion) inhibitor (EI). HIV-1 inhibitors have been optimized to prohibit the virion production based on two mechanisms of action: competitive inhibition (drug molecules that outcompete natural substrates, e.g. PI, NRTI and IN drug classes) and non-competitive inhibition (drug molecules bind at the non-active sites and induce conformational changes to prevent the substrate binding. e.g. NNRTIs and CCR5-antagonists). In clinical practice, a highly active antiretroviral therapy (HAART) for efficient HIV treatment usually combines three or more anti-HIV drugs. The HIV/AIDS guidelines in 2014 suggest the antiviral regimen to treat naïve patients, consisting of two NRTIs in combination with a third active drug from one of three classes: an NNRTI, a PI boosted with ritonavir or an integrase inhibitor (http://aidsinfo.nih.gov/guidelines). **Figure 1.10** summarizes these 5 HIV drug classes.

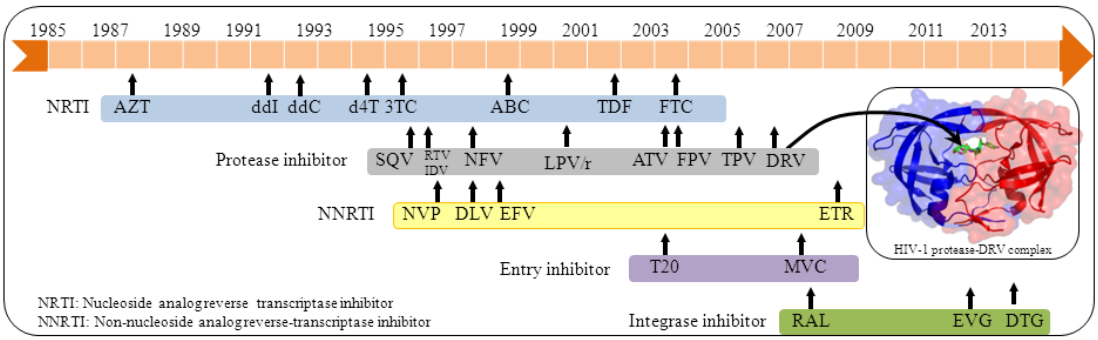


Figure 1.10: Overview of anti-HIV inhibitors approved by FDA. Five drug classes are indicated by colored bars with annotated inhibitors (see **Table 1.2**). The structure of protease-DRV complex is shown on right (PDB: 3OXW).

NRTI: AZT (Zidovudine, Retrovir) was the first anti-HIV inhibitor approved by FDA in 1987, and it was developed as an anti-cancer inhibitor during 1960s. AZT is a nucleoside analog reverse-transcriptase inhibitor (NRTI) which targets the active site of reverse transcriptase. The discovery of AZT therapy was a breakthrough in the HIV/AIDS treatment, which has significantly prevented HIV transmission and provided a promising clinical and immunologic improvement [269]. In the past two decades, other potent NRTIs have also been developed such as abacavir (ABC), emtricitabine (FTC), didanosine (ddI), lamivudine (3TC), stavudine (d4T) and

tenofovir (TDF) (**Table 1.2**). Subscribed commonly in the first-line HIV therapy, NRTIs usually have a low genetic barrier with a high risk of developing drug resistance mutations. The structural basis of HIV-1 resistance to AZT suggests that RT mutations enhance the ATP-mediated excision of AZT monophosphate from the 3' end of the DNA primer [270].

Table 1.2: Summary of FDA-approved HIV drugs (http://aidsinfo.nih.gov/).

Drug class	Name	Abbreviation	Approved date	Brand name
NRTI	Zidovudine	AZT	Mar. 1987	Retrovir
	Didanosine	ddI	Oct. 1991	Videx
	Zalcitabine	ddC	June 1992	Hivid
	Stavudine	d4T	June 1994	Zerit
	Lamivudine	3TC	Nov. 1995	Epivir
	Abacavir	ABC	Dec. 1998	Ziagen
	Tenofovir	TDF	Oct. 2001	Retrovir
	Emtricitabine	FTC	July 2003	Gilead Sciences
NNRTI	Nevirapine	NVP	June 1996	Viramune
	Delavirdine	DLV	April 1997	Rescriptor
	Efavirenz	EFV	Sept. 1998	Efavirenz
	Etravirine	ETR	Jan. 2008	Intelence
Protease inhibitor	Saquinavir	SQV	Dec. 1995	Invirase
	Ritonavir	RTV	Mar. 1996	Norvir
	Indinavir	IDV	Mar.1996	Crixivan
	Nelfinavir	NFV	Mar. 1997	Viracept
	Lopinavir	LPV	Sept. 2000	Kaletra
	Atazanavir	ATV	June 2003	Reyataz
	Fosamprenavir	FPV	Oct. 2003	Lexia
	Tipranavir	TPV	June 2005	Aptivus
	Darunavir	DRV	June 2006	Prezista
Integrase inhibitor	Raltegravir	RAL	Oct. 2007	Isentress
	Elvitegravir	EVG	Aug. 2012	Stribild
	Dolutegravir	DTG	Aug. 2013	Tivicay
Entry inhibitor	Enfuvirtide	T20	Mar. 2003	Fuzeon
	Maraviroc	MVC	Aug. 2007	Selzentry

NNRTI: NNRTIs were discovered in the late 1980s and the first NNRTI inhibitor, nevirapine, was approved in 1996 [271]. Acting as non-competitive inhibitors, NNRTIs bind with the non-cleavage site of RT to induce conformation changes, which in turn affect the catalytic activity of RT and interrupt the viral replication [272]. NNRTIs are HIV-1 specific and HIV-2 is naturally resistant to all NNRTIs [273]. When comparing RT structures between HIV-1 and HIV-2, the difference was found at both conserved and non-conserved positions in the NNRTI binding pocket (positions: 101, 106, 138, 181, 188 and 190) [274]. As of January 2014, four NNRTIs have been approved: efavirenz (EFV), etravirine (ETR), delavirdine (DLV) and nevirapine (NVP). As the potency of NNRTIs can be diminished by drug resistance mutations, novel NNRTIs have been developed and several of them are currently under clinical trials [275].

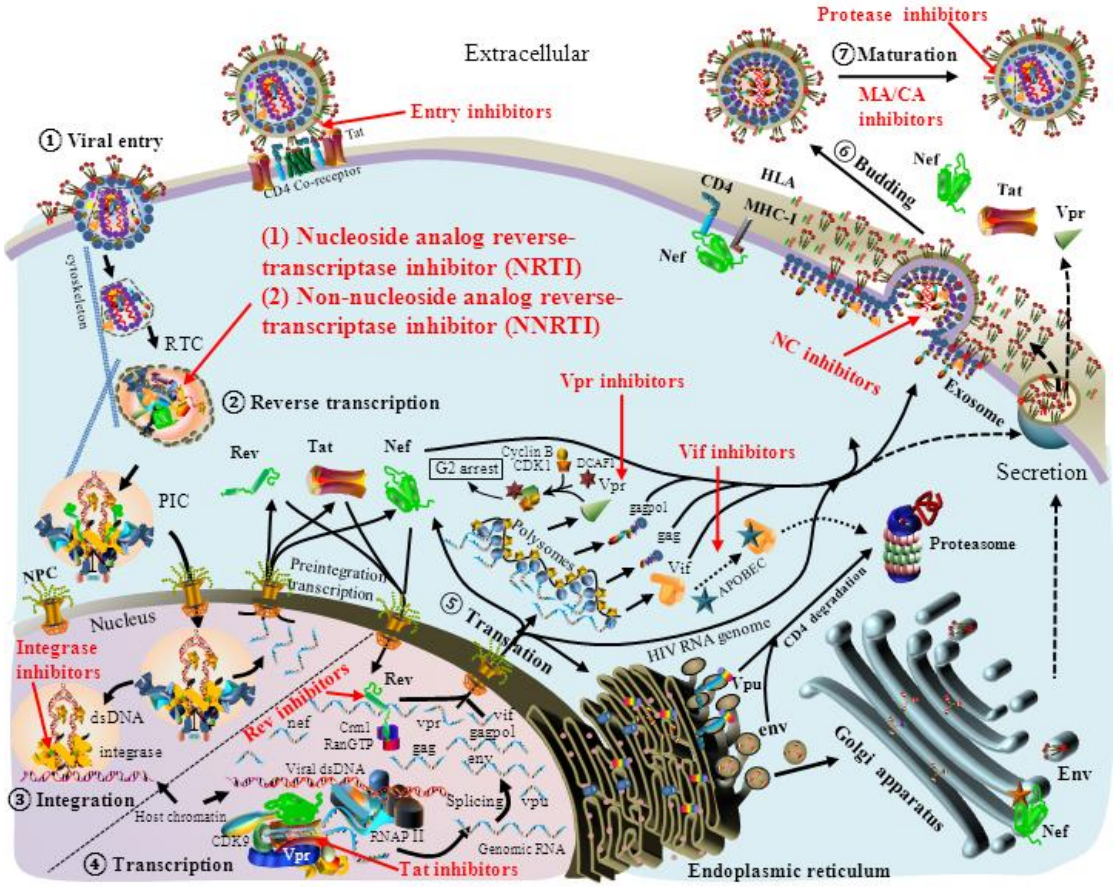


Figure 1.11: Overview of anti-HIV inhibitors in the HIV-1 life cycle. HIV inhibitors have been annotated in the HIV-1 life cycle. Descriptions and symbolic representations are described in **Figure 1.2**. Note that protein shapes and sizes are not to scale.

Protease inhibitor: In 1995, Saquinavir was approved as the first protease inhibitor, marking the start of an era for this new class of anti-HIV inhibitors. Protease inhibitors preventing viral maturation by targeting the protease cleavage site and competing with the protease substrates, namely Gag and GagPol polyproteins [12]. To date, 10 protease inhibitors have been approved by FDA: Saquinavir (SQV), Ritonavir (RTV), Indinavir (IDV), Nelfinavir (NFV), Lopinavir (LPV), Atazanavir (ATV), Fosamprenavir (FPV), Tipranavir (TPV) and Darunavir (DRV) (**Table 1.2**). Most of these PIs are prescribed along with a low dose of RTV because RTV acts as a booster for improving the bioavailability and half-life of other PIs [12]. Most clinical benefit has been shown when the protease inhibitors are considered as part of HAART. For instance, HIV RNA plasma levels were dramatically reduced when indinavir was combined with zidovudine and lamivudine in the majority of patients [12]. Yet, emerging protease drug resistance mutations challenge the potency of PIs [276]. Primary and secondary resistance mutations in protease have been documented by various clinical and experimental studies [12, 276, 277]. PI-associated mutations have been reported in the viral protease and the protease substrate Gag (Chapter 2). As an alternative mechanism for HIV to escape PI selective pressure, Gag mutations can be selected to alter structural confirmation of Gag to interact with the protease substrate-binding cleft (Chapter 4).

Integrase inhibitor: In 2007, Raltegravir was approved as the first integrase inhibitor. In the class of integrase inhibitors, raltegravir, dolutegravir and elvitegravir have become key components of anti-HIV therapy [278]. To prohibit the strand transfer reaction, these Integrase inhibitors can compete with host dsDNA to bind with the catalytic core domain of Integrase during viral integration. Compared with other anti-HIV drug classes, Integrase inhibitors have shown a good tolerability, a high safety profile and an absence of significant drug interactions [279]. Both raltegravir and elvitegravir have relatively low genetic barriers to drug resistance development, while it is not the case for dolutegravir [279]. Resistance mutations (Y143, Q148, N155) were consistently reported in all three Integrase inhibitors [278]. A better virologicval outcome of Integrase inhibitors was not significantly observed compared to protease inhibitors [279]. To develop new Integrase inhibitors, a promising class of Integrase inhibitors has been developed to target the Integrase multimerization and the

interactions between Integrase and LEDGF/p75 [280]. Several integrase inhibitors with promising antiviral activities have undergone in clinical trials [281].

Entry (fusion) inhibitor: Two entry inhibitors, Maraviroc (Selzentry, Celsentri) and Enfuvirtide (T20, Fuzeon), have been approved and many entry inhibitors have reached advanced stages of clinical trials [282]. Maraviroc is the first FDA-approved chemokine receptor antagonist or CCR5 inhibitor which targets the chemokine receptor CCR5 on the surface of CD4+ cells and macrophages [283]. CCR5 antagonists can exhibit a potent inhibition of viral replication across different HIV-1 strains. However, a significant concern was raised by the possibility that CCR5 antagonists would accelerate the disease progression by promoting the emergence of viruses with chemokine receptor CXCR4 [282, 283]. The concept of co-administration of CCR5 and CXCR4 antagonists has been halted due to the limited development of CXCR4 antagonists [282]. Enfuvirtide is a synthetic oligopeptide derived from the second helix domain (HR-2) of HIV-1 GP41. Enfuvirtide mimicks the HR-2 helix to block viral entry by preventing the interaction between the HR-1 and HR-2 helices [284]. While Enfuvirtide has a high drug efficacy with the minimal systemic toxicity, the subcutaneous administration and high cost have limited its long-term employment [282]. Following the success of Enfuvirtide, the second and third generations of GP41-derived peptide inhibitors (e.g. T1249) have been developed with promising antiviral activities in clinical and experimental studies [282, 284].

1.6 HIV functional cure

In the past three decades, many researchers have tried to develop a cure for HIV [285]. One of the underlying challenges is the persistence of a competent replication pool of HIV in the resting CD4 T cells, which makes complete HIV eradication difficult [285]. HIV in latently infected cells is not dramatically affected by the intensification of current antiretroviral therapy [285, 286]. Recent HIV studies have reported possible treatments for reaching “functional cure”, which is defined as interventions to keep the viral load at a low or undetectable level and ensure no disease progression in the absence of anti-HIV drugs. The major benefit of stepping off anti-HIV therapies is the avoidance of drug induced side effects (e.g. bone demineralization, kidney failure, etc.) [287]. Although the functional cure has been investigated for many years, only a

small proportion of HIV infected patients have been reported to maintain viral suppression in the absence of antiviral treatments [288].

News broke in March 2013 about the functional cure of a Mississippi baby has received much attention in the HIV community. It was announced that an HIV-infected baby from rural Mississippi whose medicine was stopped at 18 months of age has been living for a year without any detectable viral RNA [287]. On July 10, 2014, it was found that HIV rebound in the infected baby after antiretroviral therapy stopped for 27 months (http://www.nature.com/news/hiv-rebound-dashes-hope-of-mississippi-baby-cure-1.15535). This dashed the hope to treat infants with hard antiviral treatments for HIV eradication.

The Berlin patient is another example of a patient who is widely believed to have been cured of HIV-1 infection [286]. The 40-year-old man was infected with HIV-1 and developed acute myeloid leukaemia. In February 2007, he received a bone-marrow transplant from a donor who bears a homozygous mutation (△32) in the CCR5 chemokine receptor. This △32 mutation can naturally render the donor cells highly resistant to infections of most HIV-1 strains [289]. The antiviral treatment of this patient was stopped on the day of the first transplantation. In March 2008, he received a second transplantation with CCR5△32 stem cells from the same donor because of the relapse of acute myeloid leukemia. Surprisingly, the undetectable HIV viraemia has remained ever since [289]. Intensive efforts to identify residual HIV-1 from liver and brain showed undetectable or barely detectable level of HIV-1 DNA or RNA [289]. Due to the high cost and risk, the bone marrow transplantation is still not a solution that could be implemented to treat HIV worldwide. Yet, this case study may indicate the most relevant factors that reduce or eliminate HIV latency, which could probably provide safe and suitable alternatives in the coming years [286].

1.7 Rationale and objectives of the study

As of August 2014, a curative HIV drug or preventive vaccine remains elusive. It is also known that HIV drug resistance can impair the efficiency of all FDA-approved HIV inhibitors and challenges the development of novel inhibitors. A deep understanding of HIV genome-wide diversity, interaction and coevolution may

provide insights on the development of novel HIV inhibitors and vaccines. For instance, HIV-1 replication can be successfully blocked by targeting *gag* gene products, offering a promising strategy for new drug classes that complement contemporary HIV-1 treatment [290]. However, HIV-1 natural diversity is known to affect resistance pathways and therapy effectiveness. Natural polymorphisms are also associated with lack of response in clinical trials evaluating novel inhibitors such as Bevirimat [291]. The impact of sequence variability on drug binding sites has not been fully understood. It is therefore of interest to investigate HIV genome-wide diversity, shedding light on the drug resistance and novel drug design.

Regarding the HIV genome-wide interaction, HIV-1 proteins can interact with each other and with human proteins (see Section 1.4). This information is valuable for HIV drug design. For instance, Maraviroc is the first FDA-approved chemokine receptor antagonist or CCR5 inhibitor that targets the protein interaction between GP120 and CCR5 [283]. Bear in mind that a large number of HIV-human protein interactions have been reported recently, potential drug targets may be mined through rational drug design and a detailed mapping of HIV genome-wide protein interactions may provide guidelines for the development of novel anti-HIV inhibitors.

Coevolution in the HIV genome impacts different aspects of the HIV life cycle. The clinical relevance of HIV genome-wide coevolution is mostly involved with drug resistance. Many clinical cohort studies have shown that drug resistance mutations in drug target proteins are associated with treatment failure of HIV treatments. Recent discoveries also suggest that mutations in non-drug-target regions are also associated with treatment failure due to the genome-wide coevolution [260, 292]. For instance, amino acid substitutions in HIV Gag proteins have been recorded to compensate for the loss of binding affinity of protease mutants with the substrate Gag [293-295]. A recent study has shown that mutations at the GP41 cytoplasmic tail can confer PI drug resistance in two HIV-infected patients [260]. As indicated by Section 1.4.7, other genome-wide coevolution in the HIV genome may create new mechanisms for HIV to escape drug selective pressure.

Recent *in vitro* and *in vivo* studies have continuously reported new experimental Gag inhibitors with promising antiviral activity. While the impact of sequence variability on drug binding sites is warranted but largely lacking, the aim of Chapter 2 is to

investigate natural variations of drug binding sites across major HIV subtypes and CRFs. Previous studies that reported findings on HIV genomic diversity were limited to reference genomes or small cohorts of less than 100 patients, largely infected with a single subtype. A genome-wide analysis of HIV diversity characterizing the HIV pandemic in large-scale patient populations is still lacking. Recent advances in whole-genome sequencing, protein structure crystallization and establishment of large public databases provide large-scale data to gain new perspectives on HIV genome-wide diversity. For these reasons, the objectives of Chapter 3 were set to report the genome-wide diversity in patient populations and to identify driving factors that impact the HIV genome-wide diversity.

Detection of HIV-1 coevolution has been shown to be essential for the understanding of drug resistance, functional interactions and evolutionary pathways in HIV-1 proteins. Recent HIV-1 clinical studies indicated that treatment failure of protease inhibitors (PIs) was associated with coevolution between protease and Gag – a substrate of protease during the protease-mediated proteolytic process. However, it has not been reported which amino acid changes in Gag have been associated with drug resistance of protease inhibitors across different subtypes. For this reason, the aim of Chapter 3 is to answer this curious question. To provide standard methods for the coevolution analysis, most studies have analyzed HIV-1 protein coevolution using a single sequence-based method. However, a systematic comparison of these sequence-based methods has not been investigated, neither the possible advantages of integrating different methods to improve the prediction of HIV-1 coevolution. This research gap becomes the major focus of our study in Chapter 5. Moreover, the coevolution between the full-length Gag and protease remains unclear. By applying our method in Chapter 5, Chapter 6 wants to investigate the coevolution between the full-length Gag and protease residues to explore this new mechanism for HIV-1 viral escapes under drug selective pressure. In Chapter 7, our research turns to the study of polytree graphical models, which have been widely applied in various research fields. Previous studies have investigated the polytree graphical models under the framework of Bayesian networks – well-known probabilistic graphical models in statistics and machine learning. However, how to model the latent variables in polytree models has not been reported. The motivation of our Chapter 7 is to extend polytree models under the framework of Bayesian networks into ancestral polytree graphical models.

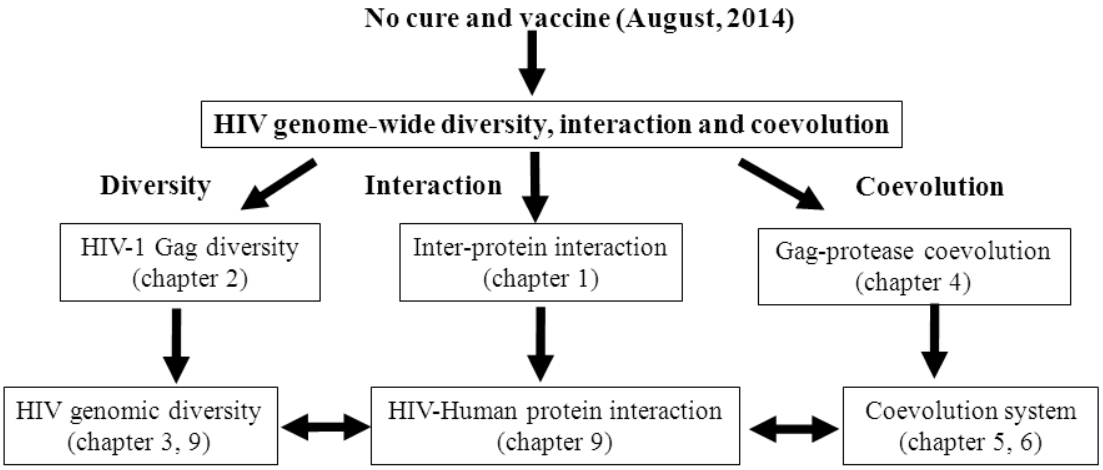


Figure 1.12: Structure of this Phd thesis.

Based on the research questions described above, this Phd thesis investigates HIV genome-wide diversity, interaction and coevolution (**Figure 1.12**). To achieve this goal, our study in Chapter 2 begins with the investigation of natural variability in HIV-1 Gag, for the reason that a great number of inhibitors have been recently developed to target Gag. Reporting natural variability in drug binding sites can hopefully improve experimental Gag inhibitors. Based on similar ideas, our study in Chapter 3 extends to the investigation of the genetic variability in the HIV full-length genome and examines driving factors that shape HIV genome-wide diversity. As one can see it, Chapter 2 and 3 focus on exploring HIV genome-wide diversity and its impact on HIV inhibitors. Afterwards, our research interests are oriented towards the understanding of HIV-1 genome-wide coevolution. Specifically, Chapter 4 used the clinical data of our Leuven patient cohorts to report the PI-associated Gag emerging during the treatment of protease inhibitors. This chapter provides the clinical background of Gag-protease coevolution in HIV-1 infected patients who received PI-based treatments. Subsequently, our investigation proposes an ensemble coevolution method (Chapter 5) and applies this method to model the Gag-protease coevolution networks (Chapter 6). Chapter 5 and 6 are sister projects that work on the computational methodology and the biological application, respectively. Our interests in Chapter 7 aim at creating new probabilistic graphical models to understand the drug resistance and protein signaling pathways, for our interests to model genome-wide associations in following studies. In the appendix, we provide detailed information on HIV genome-wide protein interactions and natural polymorphisms. More specifically, the appendix visualizes the amino acid distribution and summarizes

the literature datasets of HIV-human protein interaction in the HIV full-length genome.

Overall, we hope to provide new insights by investigating the following specific objectives: (1) investigate the sequence diversity of HIV-1 Gag, which encodes key HIV-1 structural proteins in the HIV-1 genome (Chapter 2); (2) investigate the HIV genome-wide diversity (Chapter 3); (3) evaluate the impact of HIV genomic diversity in drug resistance by analyzing HIV-1 Gag substitutions that were associated with protease drug resistance (Chapter 4); (4) design an ensemble coevolution system that integrates sequence-based statistical methods to model HIV intra- and inter-protein coevolution (Chapter 5); (5) use our ensemble coevolution system to model the networks of HIV-1 Gag-protease coevolution, which is a key component in HIV-1 genome-wide coevolution (Chapter 6); (6) propose ancestral graphical models, which allow for fast learning of large-scale interaction networks (Chapter 7). The strength and weakness of my studies, as well as their future perspectives, is addressed in the final discussion and conclusion chapter (Chapter 8).

1.8 References

1. Greene WC. A history of AIDS: looking back to see ahead. *Eur J Immunol* 2007,**37 Suppl 1**:S94-102.
2. Sharp PM, Hahn BH. Origins of HIV and the AIDS pandemic. *Cold Spring Harb Perspect Med* 2011,**1**:a006841.
3. Hemelaar J. The origin and diversity of the HIV-1 pandemic. *Trends Mol Med* 2012,**18**:182-192.
4. Hemelaar J, Gouws E, Ghys PD, Osmanov S, Isolation W-UNfH, Characterisation. Global trends in molecular epidemiology of HIV-1 during 2000-2007. *AIDS* 2011,**25**:679-689.
5. de Sousa JD, Muller V, Lemey P, Vandamme AM. High GUD incidence in the early 20 century created a particularly permissive time window for the origin and initial spread of epidemic HIV strains. *PLoS One* 2010,**5**:e9936.
6. Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, Santiago ML, *et al.* Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* 2006,**313**:523-526.
7. Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, Bunce M, *et al.* Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* 2008,**455**:661-664.
8. Abecasis AB, Vandamme AM, Lemey P. Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution. *J Virol* 2009,**83**:12917-12924.
9. Frankel AD, Young JA. HIV-1: fifteen proteins and an RNA. *Annu Rev Biochem* 1998,**67**:1-25.
10. Hill CP, Worthylake D, Bancroft DP, Christensen AM, Sundquist WI. Crystal structures of the trimeric human immunodeficiency virus type 1 matrix protein: implications for membrane association and assembly. *Proc Natl Acad Sci U S A* 1996,**93**:3099-3104.
11. Pornillos O, Ganser-Pornillos BK, Kelly BN, Hua Y, Whitby FG, Stout CD, *et al.* X-ray structures of the hexameric building block of the HIV capsid. *Cell* 2009,**137**:1282-1292.
12. Wensing AM, van Maarseveen NM, Nijhuis M. Fifteen years of HIV Protease Inhibitors: raising the barrier to resistance. *Antiviral Res* 2010,**85**:59-74.
13. Cihlar T, Ray AS. Nucleoside and nucleotide HIV reverse transcriptase inhibitors: 25 years after zidovudine. *Antiviral research* 2010,**85**:39-58.

14. Wang JY, Ling H, Yang W, Craigie R. Structure of a two-domain fragment of HIV-1 integrase: implications for domain organization in the intact protein. *EMBO J* 2001,**20**:7333-7343.
15. Auclair JR, Green KM, Shandilya S, Evans JE, Somasundaran M, Schiffer CA. Mass spectrometry analysis of HIV-1 Vif reveals an increase in ordered structure upon oligomerization in regions necessary for viral infectivity. *Proteins* 2007,**69**:270-284.
16. Frankel AD, Bredt DS, Pabo CO. Tat protein from human immunodeficiency virus forms a metal-linked dimer. *Science* 1988,**240**:70-73.
17. Daugherty MD, Liu B, Frankel AD. Structural basis for cooperative RNA binding and export complex assembly by HIV Rev. *Nat Struct Mol Biol* 2010,**17**:1337-1342.
18. Julien JP, Cupo A, Sok D, Stanfield RL, Lyumkis D, Deller MC, *et al.* Crystal structure of a soluble cleaved HIV-1 envelope trimer. *Science* 2013,**342**:1477-1483.
19. Poe JA, Smithgall TE. HIV-1 Nef dimerization is required for Nef-mediated receptor downregulation and viral replication. *J Mol Biol* 2009,**394**:329-342.
20. Waheed AA, Freed EO. HIV type 1 Gag as a target for antiviral therapy. *AIDS Res Hum Retroviruses* 2012,**28**:54-75.
21. Bell NM, Lever AM. HIV Gag polyprotein: processing and early viral particle assembly. *Trends Microbiol* 2013,**21**:136-144.
22. Zhao G, Perilla JR, Yufenyuy EL, Meng X, Chen B, Ning J, *et al.* Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* 2013,**497**:643-646.
23. Luban J. TRIM5 and the Regulation of HIV-1 Infectivity. *Mol Biol Int* 2012,**2012**:426840.
24. Mougel M, Houzet L, Darlix JL. When is it time for reverse transcription to start and go? *Retrovirology* 2009,**6**:24.
25. Darlix JL, Godet J, Ivanyi-Nagy R, Fosse P, Mauffret O, Mely Y. Flexible nature and specific functions of the HIV-1 nucleocapsid protein. *J Mol Biol* 2011,**410**:565-581.
26. Thomas JA, Gorelick RJ. Nucleocapsid protein function in early infection processes. *Virus Res* 2008,**134**:39-63.
27. Didierlaurent L, Racine PJ, Houzet L, Chamontin C, Berkhout B, Mougel M. Role of HIV-1 RNA and protein determinants for the selective packaging of spliced and unspliced viral RNA and host U6 and 7SL RNA in virus particles. *Nucleic Acids Res* 2011,**39**:8915-8927.
28. Sette P, Dussupt V, Bouamr F. Identification of the HIV-1 NC binding interface in Alix Bro1 reveals a role for RNA. *J Virol* 2012,**86**:11608-11615.
29. Xue B, Mizianty MJ, Kurgan L, Uversky VN. Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. *Cell Mol Life Sci* 2012,**69**:1211-1259.
30. Morita E, Sundquist WI. Retrovirus budding. *Annu Rev Cell Dev Biol* 2004,**20**:395-425.
31. Zybarth G, Carter C. Domains upstream of the protease (PR) in human immunodeficiency virus type 1 Gag-Pol influence PR autoprocessing. *J Virol* 1995,**69**:3878-3884.
32. Binley JM, Sanders RW, Master A, Cayanan CS, Wiley CL, Schiffner L, *et al.* Enhancing the proteolytic maturation of human immunodeficiency virus type 1 envelope glycoproteins. *J Virol* 2002,**76**:2606-2616.
33. Zhu K, Dobard C, Chow SA. Requirement for integrase during reverse transcription of human immunodeficiency virus type 1 and the effect of cysteine mutations of integrase on its interactions with reverse transcriptase. *J Virol* 2004,**78**:5045-5055.
34. Engelman A, Cherepanov P. The structural biology of HIV-1: mechanistic and therapeutic insights. *Nat Rev Microbiol* 2012,**10**:279-290.
35. Caffrey M. HIV envelope: challenges and opportunities for development of entry inhibitors. *Trends Microbiol* 2011,**19**:191-197.
36. Walker LM, Huber M, Doores KJ, Falkowska E, Pejchal R, Julien JP, *et al.* Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature* 2011,**477**:466-470.
37. Wu X, Yang ZY, Li Y, Hogerkorp CM, Schief WR, Seaman MS, *et al.* Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science* 2010,**329**:856-861.
38. Walker LM, Phogat SK, Chan-Hui PY, Wagner D, Phung P, Goss JL, *et al.* Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. *Science* 2009,**326**:285-289.
39. Postler TS, Desrosiers RC. The tale of the long tail: the cytoplasmic domain of HIV-1 gp41. *J Virol* 2013,**87**:2-15.

40. Huang J, Ofek G, Laub L, Louder MK, Doria-Rose NA, Longo NS, *et al.* Broad and potent neutralization of HIV-1 by a gp41-specific human antibody. *Nature* 2012,**491**:406-412.
41. Henriet S, Mercenne G, Bernacchi S, Paillart JC, Marquet R. Tumultuous relationship between the human immunodeficiency virus type 1 viral infectivity factor (Vif) and the human APOBEC-3G and APOBEC-3F restriction factors. *Microbiol Mol Biol Rev* 2009,**73**:211-232.
42. Jager S, Kim DY, Hultquist JF, Shindo K, LaRue RS, Kwon E, *et al.* Vif hijacks CBF-beta to degrade APOBEC3G and promote HIV-1 infection. *Nature* 2012,**481**:371-375.
43. Harris RS, Liddament MT. Retroviral restriction by APOBEC proteins. *Nat Rev Immunol* 2004,**4**:868-877.
44. Zhao RY, Li G, Bukrinsky MI. Vpr-host interactions during HIV-1 viral life cycle. *J Neuroimmune Pharmacol* 2011,**6**:216-229.
45. Cohen EA, Dehni G, Sodroski JG, Haseltine WA. Human immunodeficiency virus vpr product is a virion-associated regulatory protein. *J Virol* 1990,**64**:3097-3099.
46. Dube M, Bego MG, Paquay C, Cohen EA. Modulation of HIV-1-host interaction: role of the Vpu accessory protein. *Retrovirology* 2010,**7**:114.
47. Giroud C, Chazal N, Briant L. Cellular kinases incorporated into HIV-1 particles: passive or active passengers? *Retrovirology* 2011,**8**:71.
48. Zhu C, Gao W, Zhao K, Qin X, Zhang Y, Peng X, *et al.* Structural insight into dGTP-dependent activation of tetrameric SAMHD1 deoxynucleoside triphosphate triphosphohydrolase. *Nat Commun* 2013,**4**:2722.
49. Fujita M, Otsuka M, Nomaguchi M, Adachi A. Multifaceted activity of HIV Vpr/Vpx proteins: the current view of their virological functions. *Rev Med Virol* 2010,**20**:68-76.
50. Ayinde D, Maudet C, Transy C, Margottin-Goguet F. Limelight on two HIV/SIV accessory proteins in macrophage infection: is Vpx overshadowing Vpr? *Retrovirology* 2010,**7**:35.
51. Fujita M, Otsuka M, Miyoshi M, Khamsri B, Nomaguchi M, Adachi A. Vpx is critical for reverse transcription of the human immunodeficiency virus type 2 genome in macrophages. *J Virol* 2008,**82**:7752-7756.
52. Fernandes J, Jayaraman B, Frankel A. The HIV-1 Rev response element: an RNA scaffold that directs the cooperative assembly of a homo-oligomeric ribonucleoprotein complex. *RNA Biol* 2012,**9**:6-11.
53. Jeang K-T. Multi-Faceted Post-Transcriptional Functions of HIV-1 Rev. *Biology* 2012,**1**:165-174.
54. Strebel K. Virus-host interactions: role of HIV proteins Vif, Tat, and Rev. *AIDS* 2003,**17 Suppl 4**:S25-34.
55. Brady J, Kashanchi F. Tat gets the "green" light on transcription initiation. *Retrovirology* 2005,**2**:69.
56. Laguette N, Bregnard C, Benichou S, Basmaciogullari S. Human immunodeficiency virus (HIV) type-1, HIV-2 and simian immunodeficiency virus Nef proteins. *Mol Aspects Med* 2010,**31**:418-433.
57. Rambaut A, Posada D, Crandall KA, Holmes EC. The causes and consequences of HIV evolution. *Nat Rev Genet* 2004,**5**:52-61.
58. Neher RA, Leitner T. Recombination rate and selection strength in HIV intra-patient evolution. *PLoS Comput Biol* 2010,**6**:e1000660.
59. Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, *et al.* Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J Virol* 1999,**73**:10489-10502.
60. Jung A, Maier R, Vartanian JP, Bocharov G, Jung V, Fischer U, *et al.* Recombination: Multiply infected spleen cells in HIV patients. *Nature* 2002,**418**:144.
61. Castro-Nallar E, Perez-Losada M, Burton GF, Crandall KA. The evolution of HIV: inferences using phylogenetics. *Mol Phylogenet Evol* 2012,**62**:777-792.
62. Marsden MD, Zack JA. Eradication of HIV: current challenges and new directions. *J Antimicrob Chemother* 2009,**63**:7-10.
63. Dahl V, Josefsson L, Palmer S. HIV reservoirs, latency, and reactivation: prospects for eradication. *Antiviral Res* 2010,**85**:286-294.
64. Lamers SL, Salemi M, Galligan DC, Morris A, Gray R, Fogel G, *et al.* Human immunodeficiency virus-1 evolutionary patterns associated with pathogenic processes in the brain. *J Neurovirol* 2010,**16**:230-241.
65. Salemi M, Lamers SL, Yu S, de Oliveira T, Fitch WM, McGrath MS. Phylodynamic analysis of human immunodeficiency virus type 1 in distinct brain compartments provides a model for the neuropathogenesis of AIDS. *J Virol* 2005,**79**:11343-11352.

66. Gonzalez-Perez MP, O'Connell O, Lin R, Sullivan WM, Bell J, Simmonds P, *et al.* Independent evolution of macrophage-tropism and increased charge between HIV-1 R5 envelopes present in brain and immune tissue. *Retrovirology* 2012,**9**:20.
67. Holman AG, Mefford ME, O'Connor N, Gabuzda D. HIVBrainSeqDB: a database of annotated HIV envelope sequences from brain and other anatomical sites. *AIDS Res Ther* 2010,**7**:43.
68. Thomas ER, Dunfee RL, Stanton J, Bogdan D, Taylor J, Kunstman K, *et al.* Macrophage entry mediated by HIV Envs from brain and lymphoid tissues is determined by the capacity to use low CD4 levels and overall efficiency of fusion. *Virology* 2007,**360**:105-119.
69. Coffin JM. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science* 1995,**267**:483-489.
70. Martin N, Sattentau Q. Cell-to-cell HIV-1 spread and its implications for immune evasion. *Curr Opin HIV AIDS* 2009,**4**:143-149.
71. Murali TM, Dyer MD, Badger D, Tyler BM, Katze MG. Network-based prediction and analysis of HIV dependency factors. *PLoS Comput Biol* 2011,**7**:e1002164.
72. Bushman FD, Malani N, Fernandes J, D'Orso I, Cagney G, Diamond TL, *et al.* Host cell factors in HIV replication: meta-analysis of genome-wide studies. *PLoS Pathog* 2009,**5**:e1000437.
73. MacPherson JI, Dickerson JE, Pinney JW, Robertson DL. Patterns of HIV-1 protein interaction identify perturbed host-cellular subsystems. *PLoS Comput Biol* 2010,**6**:e1000863.
74. Konig R, Zhou Y, Elleder D, Diamond TL, Bonamy GM, Irelan JT, *et al.* Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell* 2008,**135**:49-60.
75. Jager S, Cimermancic P, Gulbahce N, Johnson JR, McGovern KE, Clarke SC, *et al.* Global landscape of HIV-human protein complexes. *Nature* 2012,**481**:365-370.
76. Fu W, Sanders-Beer BE, Katz KS, Maglott DR, Pruitt KD, Ptak RG. Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Res* 2009,**37**:D417-422.
77. Chatr-aryamontri A, Ceol A, Peluso D, Nardozza A, Panni S, Sacco F, *et al.* VirusMINT: a viral protein interaction database. *Nucleic Acids Res* 2009,**37**:D669-673.
78. Pancio HA, Ratner L. Human immunodeficiency virus type 2 Vpx-Gag interaction. *J Virol* 1998,**72**:5271-5275.
79. Warrilow D, Tachedjian G, Harrich D. Maturation of the HIV reverse transcription complex: putting the jigsaw together. *Rev Med Virol* 2009,**19**:324-337.
80. Craigie R, Bushman FD. HIV DNA Integration. *Cold Spring Harb Perspect Med* 2012,**2**:a006890.
81. Sloan RD, Wainberg MA. The role of unintegrated DNA in HIV infection. *Retrovirology* 2011,**8**:52.
82. Karn J, Stoltzfus CM. Transcriptional and posttranscriptional regulation of HIV-1 gene expression. *Cold Spring Harb Perspect Med* 2012,**2**:a006916.
83. Sundquist WI, Krausslich HG. HIV-1 Assembly, Budding, and Maturation. *Cold Spring Harb Perspect Med* 2012,**2**:a006924.
84. Sakai K, Dimas J, Lenardo MJ. The Vif and Vpr accessory proteins independently cause HIV-1-induced T cell cytopathicity and cell cycle arrest. *Proc Natl Acad Sci U S A* 2006,**103**:3369-3374.
85. Albin JS, Harris RS. Interactions of host APOBEC3 restriction factors with HIV-1 in vivo: implications for therapeutics. *Expert Rev Mol Med* 2010,**12**:e4.
86. Peterlin BM, Trono D. Hide, shield and strike back: how HIV-infected cells avoid immune eradication. *Nat Rev Immunol* 2003,**3**:97-107.
87. Wyatt R, Sodroski J. The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens. *Science* 1998,**280**:1884-1888.
88. Magadan JG, Bonifacino JS. Transmembrane domain determinants of CD4 Downregulation by HIV-1 Vpu. *J Virol* 2012,**86**:757-772.
89. Bultmann A, Muranyi W, Seed B, Haas J. Identification of two sequences in the cytoplasmic tail of the human immunodeficiency virus type 1 envelope glycoprotein that inhibit cell surface expression. *J Virol* 2001,**75**:5263-5276.
90. Hallenberger S, Bosch V, Angliker H, Shaw E, Klenk HD, Garten W. Inhibition of furin-mediated cleavage activation of HIV-1 glycoprotein gp160. *Nature* 1992,**360**:358-361.
91. Ferrucci A, Nonnemacher MR, Wigdahl B. Human immunodeficiency virus viral protein R as an extracellular protein in neuropathogenesis. *Adv Virus Res* 2011,**81**:165-199.

92. Lama J, Mangasarian A, Trono D. Cell-surface expression of CD4 reduces HIV-1 infectivity by blocking Env incorporation in a Nef- and Vpu-inhibitable manner. *Curr Biol* 1999,**9**:622-631.
93. Permanyer M, Ballana E, Este JA. Endocytosis of HIV: anything goes. *Trends Microbiol* 2010,**18**:543-551.
94. Wu L, KewalRamani VN. Dendritic-cell interactions with HIV: infection and viral dissemination. *Nat Rev Immunol* 2006,**6**:859-868.
95. Helseth E, Olshevsky U, Furman C, Sodroski J. Human immunodeficiency virus type 1 gp120 envelope glycoprotein regions important for association with the gp41 transmembrane glycoprotein. *J Virol* 1991,**65**:2119-2123.
96. York J, Nunberg JH. Role of hydrophobic residues in the central ectodomain of gp41 in maintaining the association between human immunodeficiency virus type 1 envelope glycoprotein subunits gp120 and gp41. *J Virol* 2004,**78**:4921-4926.
97. Lu M, Stoller MO, Wang S, Liu J, Fagan MB, Nunberg JH. Structural and functional analysis of interhelical interactions in the human immunodeficiency virus type 1 gp41 envelope glycoprotein by alanine-scanning mutagenesis. *J Virol* 2001,**75**:11146-11156.
98. Cao J, Bergeron L, Helseth E, Thali M, Repke H, Sodroski J. Effects of amino acid changes in the extracellular domain of the human immunodeficiency virus type 1 gp41 envelope glycoprotein. *J Virol* 1993,**67**:2747-2755.
99. Maerz AL, Drummer HE, Wilson KA, Poumbourios P. Functional analysis of the disulfide-bonded loop/chain reversal region of human immunodeficiency virus type 1 gp41 reveals a critical role in gp120-gp41 association. *J Virol* 2001,**75**:6635-6644.
100. Binley JM, Sanders RW, Clas B, Schuelke N, Master A, Guo Y, *et al.* A recombinant human immunodeficiency virus type 1 envelope glycoprotein complex stabilized by an intermolecular disulfide bond between the gp120 and gp41 subunits is an antigenic mimic of the trimeric virion-associated structure. *J Virol* 2000,**74**:627-643.
101. Yang X, Mahony E, Holm GH, Kassa A, Sodroski J. Role of the gp120 inner domain beta-sandwich in the interaction between the human immunodeficiency virus envelope glycoprotein subunits. *Virology* 2003,**313**:117-125.
102. Finzi A, Xiang SH, Pacheco B, Wang L, Haight J, Kassa A, *et al.* Topological layers in the HIV-1 gp120 inner domain regulate gp41 interaction and CD4-triggered conformational transitions. *Mol Cell* 2010,**37**:656-667.
103. Pancera M, Majeed S, Ban YE, Chen L, Huang CC, Kong L, *et al.* Structure of HIV-1 gp120 with gp41-interactive region reveals layered envelope architecture and basis of conformational mobility. *Proc Natl Acad Sci U S A* 2010,**107**:1166-1171.
104. Jacobs A, Sen J, Rong L, Caffrey M. Alanine scanning mutants of the HIV gp41 loop. *J Biol Chem* 2005,**280**:27284-27288.
105. Thali M, Furman C, Helseth E, Repke H, Sodroski J. Lack of correlation between soluble CD4-induced shedding of the human immunodeficiency virus type 1 exterior envelope glycoprotein and subsequent membrane fusion events. *J Virol* 1992,**66**:5516-5524.
106. Sen J, Jacobs A, Caffrey M. Role of the HIV gp120 conserved domain 5 in processing and viral entry. *Biochemistry* 2008,**47**:7788-7795.
107. Wang J, Sen J, Rong L, Caffrey M. Role of the HIV gp120 conserved domain 1 in processing and viral entry. *J Biol Chem* 2008,**283**:32644-32649.
108. Weng Y, Yang Z, Weiss CD. Structure-function studies of the self-assembly domain of the human immunodeficiency virus type 1 transmembrane protein gp41. *J Virol* 2000,**74**:5368-5372.
109. Gabuzda DH, Lever A, Terwilliger E, Sodroski J. Effects of deletions in the cytoplasmic domain on biological functions of human immunodeficiency virus type 1 envelope glycoproteins. *J Virol* 1992,**66**:3306-3315.
110. West JT, Weldon SK, Wyss S, Lin X, Yu Q, Thali M, *et al.* Mutation of the dominant endocytosis motif in human immunodeficiency virus type 1 gp41 can complement matrix mutations without increasing Env incorporation. *J Virol* 2002,**76**:3338-3349.
111. Davis MR, Jiang J, Zhou J, Freed EO, Aiken C. A mutation in the human immunodeficiency virus type 1 Gag protein destabilizes the interaction of the envelope protein subunits gp120 and gp41. *J Virol* 2006,**80**:2405-2417.
112. Murakami T, Freed EO. Genetic evidence for an interaction between human immunodeficiency virus type 1 matrix and alpha-helix 2 of the gp41 cytoplasmic tail. *J Virol* 2000,**74**:3548-3554.

113. Freed EO, Martin MA. Virion incorporation of envelope glycoproteins with long but not short cytoplasmic tails is blocked by specific, single amino acid substitutions in the human immunodeficiency virus type 1 matrix. *J Virol* 1995,**69**:1984-1989.
114. Bhatia AK, Campbell N, Panganiban A, Ratner L. Characterization of replication defects induced by mutations in the basic domain and C-terminus of HIV-1 matrix. *Virology* 2007,**369**:47-54.
115. Mammano F, Kondo E, Sodroski J, Bukovsky A, Gottlinger HG. Rescue of human immunodeficiency virus type 1 matrix protein mutants by envelope glycoproteins with short cytoplasmic domains. *J Virol* 1995,**69**:3824-3830.
116. Tedbury PR, Ablan SD, Freed EO. Global Rescue of Defects in HIV-1 Envelope Glycoprotein Incorporation: Implications for Matrix Structure. *PLoS Pathog* 2013,**9**:e1003739.
117. Yu X, Yuan X, Matsuda Z, Lee TH, Essex M. The matrix protein of human immunodeficiency virus type 1 is required for incorporation of viral envelope protein into mature virions. *J Virol* 1992,**66**:4966-4971.
118. Chan WE, Wang YL, Lin HH, Chen SS. Effect of extension of the cytoplasmic domain of human immunodeficiency type 1 virus transmembrane protein gp41 on virus replication. *J Virol* 2004,**78**:5157-5169.
119. Yu X, Yu QC, Lee TH, Essex M. The C terminus of human immunodeficiency virus type 1 matrix protein is involved in early steps of the virus life cycle. *J Virol* 1992,**66**:5667-5670.
120. Casella CR, Raffini LJ, Panganiban AT. Pleiotropic mutations in the HIV-1 matrix protein that affect diverse steps in replication. *Virology* 1997,**228**:294-306.
121. Marchio S, Alfano M, Primo L, Gramaglia D, Butini L, Gennero L, *et al.* Cell surface-associated Tat modulates HIV-1 infection and spreading through a specific interaction with gp120 viral envelope protein. *Blood* 2005,**105**:2802-2811.
122. Lener D, Tanchou V, Roques BP, Le Grice SF, Darlix JL. Involvement of HIV-I nucleocapsid protein in the recruitment of reverse transcriptase into nucleoprotein complexes formed in vitro. *J Biol Chem* 1998,**273**:33781-33786.
123. Druillennec S, Caneparo A, de Rocquigny H, Roques BP. Evidence of interactions between the nucleocapsid protein NCp7 and the reverse transcriptase of HIV-1. *J Biol Chem* 1999,**274**:11283-11288.
124. Cameron CE, Ghosh M, Le Grice SF, Benkovic SJ. Mutations in HIV reverse transcriptase which alter RNase H activity and decrease strand transfer efficiency are suppressed by HIV nucleocapsid protein. *Proc Natl Acad Sci U S A* 1997,**94**:6700-6705.
125. Wilkinson TA, Januszyk K, Phillips ML, Tekeste SS, Zhang M, Miller JT, *et al.* Identifying and characterizing a functional HIV-1 reverse transcriptase-binding site on integrase. *J Biol Chem* 2009,**284**:7931-7939.
126. Hehl EA, Joshi P, Kalpana GV, Prasad VR. Interaction between human immunodeficiency virus type 1 reverse transcriptase and integrase proteins. *J Virol* 2004,**78**:5056-5067.
127. Oz Gleenberg I, Goldgur Y, Hizi A. Ile178 of HIV-1 reverse transcriptase is critical for inhibiting the viral integrase. *Biochem Biophys Res Commun* 2007,**364**:48-52.
128. Kataropoulou A, Bovolenta C, Belfiore A, Trabatti S, Garbelli A, Porcellini S, *et al.* Mutational analysis of the HIV-1 auxiliary protein Vif identifies independent domains important for the physical and functional interaction with HIV-1 reverse transcriptase. *Nucleic Acids Res* 2009,**37**:3660-3669.
129. Nascimbeni M, Bouyac M, Rey F, Spire B, Clavel F. The replicative impairment of Vif-mutants of human immunodeficiency virus type 1 correlates with an overall defect in viral DNA synthesis. *J Gen Virol* 1998,**79 (Pt 8)**:1945-1950.
130. Apolloni A, Hooker CW, Mak J, Harrich D. Human immunodeficiency virus type 1 protease regulation of tat activity is essential for efficient reverse transcription and replication. *J Virol* 2003,**77**:9912-9921.
131. Harrich D, Ulich C, Garcia-Martinez LF, Gaynor RB. Tat is required for efficient HIV-1 reverse transcription. *EMBO J* 1997,**16**:1224-1235.
132. Kameoka M, Rong L, Gotte M, Liang C, Russell RS, Wainberg MA. Role for human immunodeficiency virus type 1 Tat protein in suppression of viral reverse transcriptase activity during late stages of viral replication. *J Virol* 2001,**75**:2675-2683.
133. Apolloni A, Meredith LW, Suhrbier A, Kiernan R, Harrich D. The HIV-1 Tat protein stimulates reverse transcription in vitro. *Curr HIV Res* 2007,**5**:473-483.
134. Fournier C, Cortay JC, Carbonnelle C, Ehresmann C, Marquet R, Boulanger P. The HIV-1 Nef protein enhances the affinity of reverse transcriptase for RNA in vitro. *Virus Genes* 2002,**25**:255-269.

135. Ciuffi A, Munoz M, Bleiber G, Favre M, Stutz F, Telenti A, *et al.* Interactions of processed Nef (58-206) with virion proteins of HIV type 1. *AIDS Res Hum Retroviruses* 2004,**20**:399-407.
136. Rosenbluh J, Hayouka Z, Loya S, Levin A, Armon-Omer A, Britan E, *et al.* Interaction between HIV-1 Rev and integrase proteins: a basis for the development of anti-HIV peptides. *J Biol Chem* 2007,**282**:15743-15753.
137. Levin A, Rosenbluh J, Hayouka Z, Friedler A, Loyter A. Integration of HIV-1 DNA is regulated by interplay between viral rev and cellular LEDGF/p75 proteins. *Mol Med* 2010,**16**:34-44.
138. Levin A, Hayouka Z, Helfer M, Brack-Werner R, Friedler A, Loyter A. Peptides derived from HIV-1 integrase that bind Rev stimulate viral genome integration. *PLoS One* 2009,**4**:e4155.
139. Gallay P, Swingler S, Song J, Bushman F, Trono D. HIV nuclear import is governed by the phosphotyrosine-mediated binding of matrix to the core domain of integrase. *Cell* 1995,**83**:569-576.
140. Sato A, Yoshimoto J, Isaka Y, Miki S, Suyama A, Adachi A, *et al.* Evidence for direct association of Vpr and matrix protein p17 within the HIV-1 virion. *Virology* 1996,**220**:208-212.
141. Sawaya BE, Khalili K, Gordon J, Taube R, Amini S. Cooperative interaction between HIV-1 regulatory proteins Tat and Vpr modulates transcription of the viral genome. *J Biol Chem* 2000,**275**:35209-35214.
142. Liao WH, Huang KJ, Chang YF, Wang SM, Tseng YT, Chiang CC, *et al.* Incorporation of human immunodeficiency virus type 1 reverse transcriptase into virus-like particles. *J Virol* 2007,**81**:5155-5165.
143. de Rocquigny H, Petitjean P, Tanchou V, Decimo D, Drouot L, Delaunay T, *et al.* The zinc fingers of HIV nucleocapsid protein NCp7 direct interactions with the viral regulatory protein Vpr. *J Biol Chem* 1997,**272**:30753-30759.
144. Bouyac M, Courcoul M, Bertoia G, Baudat Y, Gabuzda D, Blanc D, *et al.* Human immunodeficiency virus type 1 Vif protein binds to the Pr55Gag precursor. *J Virol* 1997,**71**:9358-9365.
145. Huvent I, Hong SS, Fournier C, Gay B, Tournier J, Carriere C, *et al.* Interaction and co-encapsidation of human immunodeficiency virus type 1 Gag and Vif recombinant proteins. *J Gen Virol* 1998,**79 (Pt 5)**:1069-1081.
146. Sova P, Volsky DJ, Wang L, Chao W. Vif is largely absent from human immunodeficiency virus type 1 mature virions and associates mainly with viral particles containing unprocessed gag. *J Virol* 2001,**75**:5504-5517.
147. Costa LJ, Zheng YH, Sabotic J, Mak J, Fackler OT, Peterlin BM. Nef binds p6* in GagPol during replication of human immunodeficiency virus type 1. *J Virol* 2004,**78**:5311-5323.
148. Schiavoni I, Trapp S, Santarcangelo AC, Piacentini V, Pugliese K, Baur A, *et al.* HIV-1 Nef enhances both membrane expression and virion incorporation of Env products. A model for the Nef-dependent increase of HIV-1 infectivity. *J Biol Chem* 2004,**279**:22996-23006.
149. Paxton W, Connor RI, Landau NR. Incorporation of Vpr into human immunodeficiency virus type 1 virions: requirement for the p6 region of gag and mutational analysis. *J Virol* 1993,**67**:7229-7237.
150. Jenkins Y, Pornillos O, Rich RL, Myszka DG, Sundquist WI, Malim MH. Biochemical analyses of the interactions between human immunodeficiency virus type 1 Vpr and p6(Gag). *J Virol* 2001,**75**:10537-10542.
151. Kondo E, Gottlinger HG. A conserved LXXLF sequence is the major determinant in p6gag required for the incorporation of human immunodeficiency virus type 1 Vpr. *J Virol* 1996,**70**:159-164.
152. Lu YL, Bennett RP, Wills JW, Gorelick R, Ratner L. A leucine triplet repeat sequence (LXX)4 in p6gag is important for Vpr incorporation into human immunodeficiency virus type 1 particles. *J Virol* 1995,**69**:6873-6879.
153. Zhu H, Jian H, Zhao LJ. Identification of the 15FRFG domain in HIV-1 Gag p6 essential for Vpr packaging into the virion. *Retrovirology* 2004,**1**:26.
154. Salgado GF, Marquant R, Vogel A, Alves ID, Feller SE, Morellet N, *et al.* Structural studies of HIV-1 Gag p6ct and its interaction with Vpr determined by solution nuclear magnetic resonance. *Biochemistry* 2009,**48**:2355-2367.
155. Ludwig C, Leiherer A, Wagner R. Importance of protease cleavage sites within and flanking human immunodeficiency virus type 1 transframe protein p6* for spatiotemporal regulation of protease activation. *J Virol* 2008,**82**:4573-4584.

156. Hutoran M, Britan E, Baraz L, Blumenzweig I, Steinitz M, Kotler M. Abrogation of Vif function by peptide derived from the N-terminal region of the human immunodeficiency virus type 1 (HIV-1) protease. *Virology* 2004,**330**:261-270.
157. Potash MJ, Bentsman G, Muir T, Krachmarov C, Sova P, Volsky DJ. Peptide inhibitors of HIV-1 protease and viral infection of peripheral blood lymphocytes based on HIV-1 Vif. *Proc Natl Acad Sci U S A* 1998,**95**:13865-13868.
158. Baraz L, Hutoran M, Blumenzweig I, Katzenellenbogen M, Friedler A, Gilon C, *et al.* Human immunodeficiency virus type 1 Vif binds the viral protease by interaction with its N-terminal region. *J Gen Virol* 2002,**83**:2225-2230.
159. Friedler A, Blumenzweig I, Baraz L, Steinitz M, Kotler M, Gilon C. Peptides derived from HIV-1 Vif: a non-substrate based novel type of HIV-1 protease inhibitors. *J Mol Biol* 1999,**287**:93-101.
160. Chan DC, Kim PS. HIV entry and its inhibition. *Cell* 1998,**93**:681-684.
161. Melikyan GB. Common principles and intermediates of viral protein-mediated fusion: the HIV-1 paradigm. *Retrovirology* 2008,**5**:111.
162. Sierra-Aragon S, Walter H. Targets for inhibition of HIV replication: entry, enzyme action, release and maturation. *Intervirology* 2012,**55**:84-97.
163. Uchil PD, Mothes W. HIV Entry Revisited. *Cell* 2009,**137**:402-404.
164. Yi L, Fang J, Isik N, Chim J, Jin T. HIV gp120-induced interaction between CD4 and CCR5 requires cholesterol-rich microenvironments revealed by live cell fluorescence resonance energy transfer imaging. *J Biol Chem* 2006,**281**:35446-35453.
165. Kim S, Pang HB, Kay MS. Peptide mimic of the HIV envelope gp120-gp41 interface. *J Mol Biol* 2008,**376**:786-797.
166. Kowalski M, Potz J, Basiripour L, Dorfman T, Goh WC, Terwilliger E, *et al.* Functional regions of the envelope glycoprotein of human immunodeficiency virus type 1. *Science* 1987,**237**:1351-1355.
167. Sen J, Yan T, Wang J, Rong L, Tao L, Caffrey M. Alanine scanning mutagenesis of HIV-1 gp41 heptad repeat 1: insight into the gp120-gp41 interaction. *Biochemistry* 2010,**49**:5057-5065.
168. Abrahamyan LG, Mkrtchyan SR, Binley J, Lu M, Melikyan GB, Cohen FS. The cytoplasmic tail slows the folding of human immunodeficiency virus type 1 Env from a late prebundle configuration into the six-helix bundle. *J Virol* 2005,**79**:106-115.
169. Murakami T, Freed EO. The long cytoplasmic tail of gp41 is required in a cell type-dependent manner for HIV-1 envelope glycoprotein incorporation into virions. *Proc Natl Acad Sci U S A* 2000,**97**:343-348.
170. Jiang J, Aiken C. Maturation of the viral core enhances the fusion of HIV-1 particles with primary human T cells and monocyte-derived macrophages. *Virology* 2006,**346**:460-468.
171. Dorfman T, Mammano F, Haseltine WA, Gottlinger HG. Role of the matrix protein in the virion association of the human immunodeficiency virus type 1 envelope glycoprotein. *J Virol* 1994,**68**:1689-1696.
172. Poon S, Moscoso CG, Xing L, Kan E, Sun Y, Kolatkar PR, *et al.* Putative role of Tat-Env interaction in HIV infection. *AIDS* 2013,**27**:2345-2354.
173. Frankel AD, Pabo CO. Cellular uptake of the tat protein from human immunodeficiency virus. *Cell* 1988,**55**:1189-1193.
174. Ensoli B, Buonaguro L, Barillari G, Fiorelli V, Gendelman R, Morgan RA, *et al.* Release, uptake, and effects of extracellular human immunodeficiency virus type 1 Tat protein on cell growth and viral transactivation. *J Virol* 1993,**67**:277-287.
175. Albini A, Ferrini S, Benelli R, Sforzini S, Giunciuglio D, Aluigi MG, *et al.* HIV-1 Tat protein mimicry of chemokines. *Proc Natl Acad Sci U S A* 1998,**95**:13153-13158.
176. Guy B, Geist M, Dott K, Spehner D, Kieny MP, Lecocq JP. A specific inhibitor of cysteine proteases impairs a Vif-dependent modification of human immunodeficiency virus type 1 Env protein. *J Virol* 1991,**65**:1325-1331.
177. Ma XY, Sova P, Chao W, Volsky DJ. Cysteine residues in the Vif protein of human immunodeficiency virus type 1 are essential for viral infectivity. *J Virol* 1994,**68**:1714-1720.
178. Akari H, Yoshida A, Fukumori T, Adachi A. Host cell-dependent replication of HIV-1 mutants with deletions in gp41 cytoplasmic tail region is independent of the function of Vif. *Microbes Infect* 2000,**2**:1019-1023.
179. Hu WS, Hughes SH. HIV-1 reverse transcription. *Cold Spring Harb Perspect Med* 2012,**2**.
180. Zennou V, Petit C, Guetard D, Nerhbass U, Montagnier L, Charneau P. HIV-1 genome nuclear import is mediated by a central DNA flap. *Cell* 2000,**101**:173-185.

181. Basu VP, Song M, Gao L, Rigby ST, Hanson MN, Bambara RA. Strand transfer events during HIV-1 reverse transcription. *Virus Res* 2008,**134**:19-38.
182. Lai RP, Yan J, Heeney J, McClure MO, Gottlinger H, Luban J, *et al.* Nef decreases HIV-1 sensitivity to neutralizing antibodies that target the membrane-proximal external region of TMgp41. *PLoS Pathog* 2011,**7**:e1002442.
183. Arhel NJ, Souquere-Besse S, Munier S, Souque P, Guadagnini S, Rutherford S, *et al.* HIV-1 DNA Flap formation promotes uncoating of the pre-integration complex at the nuclear pore. *EMBO J* 2007,**26**:3025-3037.
184. Arhel N. Revisiting HIV-1 uncoating. *Retrovirology* 2010,**7**:96.
185. Roda RH, Balakrishnan M, Hanson MN, Wohrl BM, Le Grice SF, Roques BP, *et al.* Role of the Reverse Transcriptase, Nucleocapsid Protein, and Template Structure in the Two-step Transfer Mechanism in Retroviral Recombination. *J Biol Chem* 2003,**278**:31536-31546.
186. Tasara T, Maga G, Hottiger MO, Hubscher U. HIV-1 reverse transcriptase and integrase enzymes physically interact and inhibit each other. *FEBS Lett* 2001,**507**:39-44.
187. Herschhorn A, Oz-Gleenberg I, Hizi A. Quantitative analysis of the interactions between HIV-1 integrase and retroviral reverse transcriptases. *Biochem J* 2008,**412**:163-170.
188. Wu X, Liu H, Xiao H, Conway JA, Hehl E, Kalpana GV, *et al.* Human immunodeficiency virus type 1 integrase protein promotes reverse transcription through specific interactions with the nucleoprotein reverse transcription complex. *J Virol* 1999,**73**:2126-2135.
189. Dobard CW, Briones MS, Chow SA. Molecular mechanisms by which human immunodeficiency virus type 1 integrase stimulates the early steps of reverse transcription. *J Virol* 2007,**81**:10037-10046.
190. Barat C, Lullien V, Schatz O, Keith G, Nugeyre MT, Gruninger-Leitch F, *et al.* HIV-1 reverse transcriptase specifically interacts with the anticodon domain of its cognate primer tRNA. *EMBO J* 1989,**8**:3279-3285.
191. Kim J, Roberts A, Yuan H, Xiong Y, Anderson KS. Nucleocapsid protein annealing of a primer-template enhances (+)-strand DNA synthesis and fidelity by HIV-1 reverse transcriptase. *J Mol Biol* 2012,**415**:866-880.
192. Grohmann D, Godet J, Mely Y, Darlix JL, Restle T. HIV-1 nucleocapsid traps reverse transcriptase on nucleic acid substrates. *Biochemistry* 2008,**47**:12230-12240.
193. Simon JH, Malim MH. The human immunodeficiency virus type 1 Vif protein modulates the postpenetration stability of viral nucleoprotein complexes. *J Virol* 1996,**70**:5297-5305.
194. Dettenhofer M, Cen S, Carlson BA, Kleiman L, Yu XF. Association of human immunodeficiency virus type 1 Vif with RNA and its role in reverse transcription. *J Virol* 2000,**74**:8938-8945.
195. Dornadula G, Yang S, Pomerantz RJ, Zhang H. Partial rescue of the Vif-negative phenotype of mutant human immunodeficiency virus type 1 strains from nonpermissive cells by intravirion reverse transcription. *J Virol* 2000,**74**:2594-2602.
196. Goncalves J, Korin Y, Zack J, Gabuzda D. Role of Vif in human immunodeficiency virus type 1 reverse transcription. *J Virol* 1996,**70**:8701-8709.
197. Ulich C, Dunne A, Parry E, Hooker CW, Gaynor RB, Harrich D. Functional domains of Tat required for efficient human immunodeficiency virus type 1 reverse transcription. *J Virol* 1999,**73**:2499-2508.
198. Kameoka M, Morgan M, Binette M, Russell RS, Rong L, Guo X, *et al.* The Tat protein of human immunodeficiency virus type 1 (HIV-1) can promote placement of tRNA primer onto viral RNA and suppress later DNA polymerization in HIV-1 reverse transcription. *J Virol* 2002,**76**:3637-3645.
199. Popov S, Rexach M, Ratner L, Blobel G, Bukrinsky M. Viral protein R regulates docking of the HIV-1 preintegration complex to the nuclear pore complex. *J Biol Chem* 1998,**273**:13347-13352.
200. Gleenberg IO, Herschhorn A, Hizi A. Inhibition of the activities of reverse transcriptase and integrase of human immunodeficiency virus type-1 by peptides derived from the homologous viral protein R (Vpr). *J Mol Biol* 2007,**369**:1230-1243.
201. Stark LA, Hay RT. Human immunodeficiency virus type 1 (HIV-1) viral protein R (Vpr) interacts with Lys-tRNA synthetase: implications for priming of HIV-1 reverse transcription. *J Virol* 1998,**72**:3037-3044.
202. Nermut MV, Fassati A. Structural analyses of purified human immunodeficiency virus type 1 intracellular reverse transcription complexes. *J Virol* 2003,**77**:8196-8206.
203. Fassati A, Goff SP. Characterization of intracellular reverse transcription complexes of human immunodeficiency virus type 1. *J Virol* 2001,**75**:3626-3635.

204. Suzuki Y, Craigie R. The road to chromatin - nuclear entry of retroviruses. *Nat Rev Microbiol* 2007,**5**:187-196.
205. Poeschla EM. Integrase, LEDGF/p75 and HIV replication. *Cell Mol Life Sci* 2008,**65**:1403-1424.
206. Oz I, Avidan O, Hizi A. Inhibition of the integrases of human immunodeficiency viruses type 1 and type 2 by reverse transcriptases. *Biochem J* 2002,**361**:557-566.
207. Pommier Y, Johnson AA, Marchand C. Integrase inhibitors to treat HIV/AIDS. *Nat Rev Drug Discov* 2005,**4**:236-248.
208. Heinzinger NK, Bukinsky MI, Haggerty SA, Ragland AM, Kewalramani V, Lee MA, *et al.* The Vpr protein of human immunodeficiency virus type 1 influences nuclear localization of viral nucleic acids in nondividing host cells. *Proc Natl Acad Sci U S A* 1994,**91**:7311-7315.
209. Dahiya S, Nonnemacher MR, Wigdahl B. Deployment of the human immunodeficiency virus type 1 protein arsenal: combating the host to enhance viral transcription and providing targets for therapeutic development. *J Gen Virol* 2012,**93**:1151-1172.
210. He N, Zhou Q. New insights into the control of HIV-1 transcription: when Tat meets the 7SK snRNP and super elongation complex (SEC). *J Neuroimmune Pharmacol* 2011,**6**:260-268.
211. Poon B, Chang MA, Chen IS. Vpr is required for efficient Nef expression from unintegrated human immunodeficiency virus type 1 DNA. *J Virol* 2007,**81**:10515-10523.
212. Joseph AM, Ladha JS, Mojamdar M, Mitra D. Human immunodeficiency virus-1 Nef protein interacts with Tat and enhances HIV-1 gene expression. *FEBS Lett* 2003,**548**:37-42.
213. Wolf D, Witte V, Clark P, Blume K, Lichtenheld MG, Baur AS. HIV Nef enhances Tat-mediated viral transcription through a hnRNP-K-nucleated signaling complex. *Cell Host Microbe* 2008,**4**:398-408.
214. Witte V, Laffert B, Gintschel P, Krautkramer E, Blume K, Fackler OT, *et al.* Induction of HIV transcription by Nef involves Lck activation and protein kinase C theta raft recruitment leading to activation of ERK1/2 but not NF kappa B. *J Immunol* 2008,**181**:8425-8432.
215. Simmons A, Aluvihare V, McMichael A. Nef triggers a transcriptional program in T cells imitating single-signal T cell activation and inducing HIV virulence mediators. *Immunity* 2001,**14**:763-777.
216. Zhou C, Rana TM. A bimolecular mechanism of HIV-1 Tat protein interaction with RNA polymerase II transcription elongation complexes. *J Mol Biol* 2002,**320**:925-942.
217. Wang J, Shackelford JM, Selliah N, Shivers DK, O'Neill E, Garcia JV, *et al.* The HIV-1 Vif protein mediates degradation of Vpr and reduces Vpr-induced cell cycle arrest. *DNA Cell Biol* 2008,**27**:267-277.
218. Turelli P, Doucas V, Craig E, Mangeat B, Klages N, Evans R, *et al.* Cytoplasmic recruitment of INI1 and PML on incoming HIV preintegration complexes: interference with early steps of viral replication. *Mol Cell* 2001,**7**:1245-1254.
219. Sorin M, Yung E, Wu X, Kalpana GV. HIV-1 replication in cell lines harboring INI1/hSNF5 mutations. *Retrovirology* 2006,**3**:56.
220. Ariumi Y, Serhan F, Turelli P, Telenti A, Trono D. The integrase interactor 1 (INI1) proteins facilitate Tat-mediated human immunodeficiency virus type 1 transcription. *Retrovirology* 2006,**3**:47.
221. Pornillos O, Garrus JE, Sundquist WI. Mechanisms of enveloped RNA virus budding. *Trends Cell Biol* 2002,**12**:569-579.
222. Murakami T. Retroviral env glycoprotein trafficking and incorporation into virions. *Mol Biol Int* 2012,**2012**:682850.
223. Lu K, Heng X, Summers MF. Structural determinants and mechanism of HIV-1 genome packaging. *J Mol Biol* 2011,**410**:609-633.
224. Holmes RK, Malim MH, Bishop KN. APOBEC-mediated viral restriction: not simply editing? *Trends Biochem Sci* 2007,**32**:118-128.
225. Waheed AA, Freed EO. Lipids and membrane microdomains in HIV-1 replication. *Virus Res* 2009,**143**:162-176.
226. Weiss ER, Gottlinger H. The role of cellular factors in promoting HIV budding. *J Mol Biol* 2011,**410**:525-533.
227. Henzler T, Harmache A, Herrmann H, Spring H, Suzan M, Audoly G, *et al.* Fully functional, naturally occurring and C-terminally truncated variant human immunodeficiency virus (HIV) Vif does not bind to HIV Gag but influences intermediate filament structure. *J Gen Virol* 2001,**82**:561-573.
228. Seroude V, Audoly G, Gluschankof P, Suzan M. Viral and cellular specificities of caprine arthritis encephalitis virus Vif protein. *Virology* 2002,**292**:156-161.

229. Henriet S, Sinck L, Bec G, Gorelick RJ, Marquet R, Paillart JC. Vif is a RNA chaperone that could temporally regulate RNA dimerization and the early steps of HIV-1 reverse transcription. *Nucleic Acids Res* 2007,**35**:5141-5153.
230. Akari H, Fujita M, Kao S, Khan MA, Shehu-Xhilaga M, Adachi A, *et al.* High level expression of human immunodeficiency virus type-1 Vif inhibits viral infectivity by modulating proteolytic processing of the Gag precursor at the p2/nucleocapsid processing site. *J Biol Chem* 2004,**279**:12355-12362.
231. Ohagen A, Gabuzda D. Role of Vif in stability of the human immunodeficiency virus type 1 core. *J Virol* 2000,**74**:11055-11066.
232. Kondo E, Mammano F, Cohen EA, Gottlinger HG. The p6gag domain of human immunodeficiency virus type 1 is sufficient for the incorporation of Vpr into heterologous viral particles. *J Virol* 1995,**69**:2759-2764.
233. Bachand F, Yao XJ, Hrimech M, Rougeau N, Cohen EA. Incorporation of Vpr into human immunodeficiency virus type 1 requires a direct interaction with the p6 domain of the p55 gag precursor. *J Biol Chem* 1999,**274**:9083-9091.
234. Salgado GF, Vogel A, Marquant R, Feller SE, Bouaziz S, Alves ID. The role of membranes in the organization of HIV-1 Gag p6 and Vpr: p6 shows high affinity for membrane bilayers which substantially increases the interaction between p6 and Vpr. *J Med Chem* 2009,**52**:7157-7162.
235. Lopez-Verges S, Camus G, Blot G, Beauvoir R, Benarous R, Berlioz-Torrent C. Tail-interacting protein TIP47 is a connector between Gag and Env and is required for Env incorporation into HIV-1 virions. *Proc Natl Acad Sci U S A* 2006,**103**:14947-14952.
236. Li MS, Garcia-Asua G, Bhattacharyya U, Mascagni P, Austen BM, Roberts MM. The Vpr protein of human immunodeficiency virus type 1 binds to nucleocapsid protein p7 in vitro. *Biochem Biophys Res Commun* 1996,**218**:352-355.
237. de Rocquigny H, Caneparo A, Delaunay T, Bischerour J, Mouscadet JF, Roques BP. Interactions of the C-terminus of viral protein R with nucleic acids are modulated by its N-terminus. *Eur J Biochem* 2000,**267**:3654-3660.
238. Pizzato M, Popova E, Gottlinger HG. Nef can enhance the infectivity of receptor-pseudotyped human immunodeficiency virus type 1 particles. *J Virol* 2008,**82**:10811-10819.
239. Zhou J, Aiken C. Nef enhances human immunodeficiency virus type 1 infectivity resulting from intervirion fusion: evidence supporting a role for Nef at the virion envelope. *J Virol* 2001,**75**:5851-5859.
240. Ono T, Iwatani Y, Nishimura A, Ishimoto A, Sakai H. Functional association between the nef gene product and gag-pol region of HIV-1. *FEBS Lett* 2000,**466**:233-238.
241. Weclewicz K, Ekstrom M, Kristensson K, Garoff H. Specific interactions between retrovirus Env and Gag proteins in rat neurons. *J Virol* 1998,**72**:2832-2845.
242. Hermida-Matsumoto L, Resh MD. Localization of human immunodeficiency virus type 1 Gag and Env at the plasma membrane by confocal imaging. *J Virol* 2000,**74**:8670-8679.
243. Bhattacharya J, Repik A, Clapham PR. Gag regulates association of human immunodeficiency virus type 1 envelope with detergent-resistant membranes. *J Virol* 2006,**80**:5292-5300.
244. Shiraishi T, Misumi S, Takama M, Takahashi I, Shoji S. Myristoylation of human immunodeficiency virus type 1 gag protein is required for efficient env protein transportation to the surface of cells. *Biochem Biophys Res Commun* 2001,**282**:1201-1205.
245. Deora A, Spearman P, Ratner L. The N-terminal matrix domain of HIV-1 Gag is sufficient but not necessary for viral protein U-mediated enhancement of particle release through a membrane-targeting mechanism. *Virology* 2000,**269**:305-312.
246. Yeager M. Design of in vitro symmetric complexes and analysis by hybrid methods reveal mechanisms of HIV capsid assembly. *J Mol Biol* 2011,**410**:534-552.
247. Hill M, Tachedjian G, Mak J. The packaging and maturation of the HIV-1 Pol proteins. *Curr HIV Res* 2005,**3**:73-85.
248. Baraz L, Friedler A, Blumenzweig I, Nussinuv O, Chen N, Steinitz M, *et al.* Human immunodeficiency virus type 1 Vif-derived peptides inhibit the viral protease and arrest virus production. *FEBS Lett* 1998,**441**:419-426.
249. Adekale MA, Cane PA, McCrae MA. Changes in the Vif protein of HIV-1 associated with the development of resistance to inhibitors of viral protease. *J Med Virol* 2005,**75**:195-201.
250. Bottcher M, Grosse F. HIV-1 protease inhibits its homologous reverse transcriptase by protein-protein interaction. *Nucleic Acids Res* 1997,**25**:1709-1714.
251. Goobar-Larsson L, Larsson PT, Debouck C, Towler EM. HIV-1 RT enhances the activity of a tethered dimer of HIV-1 proteinase. *Biochem Biophys Res Commun* 1996,**220**:203-207.

252. Goobar-Larsson L, Luukkonen BG, Unge T, Schwartz S, Utter G, Strandberg B, *et al.* Enhancement of HIV-1 proteinase activity by HIV-1 reverse transcriptase. *Virology* 1995,**206**:387-394.
253. Trubey CM, Chertova E, Coren LV, Hilburn JM, Hixson CV, Nagashima K, *et al.* Quantitation of HLA class II protein incorporated into human immunodeficiency type 1 virions purified by anti-CD45 immunoaffinity depletion of microvesicles. *J Virol* 2003,**77**:12699-12709.
254. Zhu P, Liu J, Bess J, Jr., Chertova E, Lifson JD, Grise H, *et al.* Distribution and three-dimensional structure of AIDS virus envelope spikes. *Nature* 2006,**441**:847-852.
255. Zhu P, Chertova E, Bess J, Jr., Lifson JD, Arthur LO, Liu J, *et al.* Electron tomography analysis of envelope glycoprotein trimers on HIV and simian immunodeficiency virus virions. *Proc Natl Acad Sci U S A* 2003,**100**:15812-15817.
256. Fouchier RA, Simon JH, Jaffe AB, Malim MH. Human immunodeficiency virus type 1 Vif does not influence expression or virion incorporation of gag-, pol-, and env-encoded proteins. *J Virol* 1996,**70**:8263-8269.
257. Bischerour J, Tauc P, Leh H, de Rocquigny H, Roques B, Mouscadet JF. The (52-96) C-terminal domain of Vpr stimulates HIV-1 IN-mediated homologous strand transfer of mini-viral DNA. *Nucleic Acids Res* 2003,**31**:2694-2702.
258. Briones MS, Dobard CW, Chow SA. Role of human immunodeficiency virus type 1 integrase in uncoating of the viral core. *J Virol* 2010,**84**:5181-5190.
259. Dismuke DJ, Aiken C. Evidence for a functional link between uncoating of the human immunodeficiency virus type 1 core and nuclear import of the viral preintegration complex. *J Virol* 2006,**80**:3712-3720.
260. Rabi SA, Laird GM, Durand CM, Laskey S, Shan L, Bailey JR, *et al.* Multi-step inhibition explains HIV-1 protease inhibitor pharmacodynamics and resistance. *J Clin Invest* 2013,**123**:3848-3860.
261. Stephenson KE, Barouch DH. A global approach to HIV-1 vaccine development. *Immunol Rev* 2013,**254**:295-304.
262. Cohen YZ, Dolin R. Novel HIV vaccine strategies: overview and perspective. *Ther Adv Vaccines* 2013,**1**:99-112.
263. Sanou MP, De Groot AS, Murphey-Corb M, Levy JA, Yamamoto JK. HIV-1 Vaccine Trials: Evolving Concepts and Designs. *Open AIDS J* 2012,**6**:274-288.
264. Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, Kaewkungwal J, Chiu J, Paris R, *et al.* Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *N Engl J Med* 2009,**361**:2209-2220.
265. Bos R, van Duikeren S, van Hall T, Lauwen MM, Parrington M, Berinstein NL, *et al.* Characterization of antigen-specific immune responses induced by canarypox virus vaccines. *J Immunol* 2007,**179**:6115-6122.
266. Rolland M, Gilbert P. Evaluating immune correlates in HIV type 1 vaccine efficacy trials: what RV144 may provide. *AIDS Res Hum Retroviruses* 2012,**28**:400-404.
267. Mehellou Y, De Clercq E. Twenty-six years of anti-HIV drug discovery: where do we stand and where do we go? *J Med Chem* 2010,**53**:521-538.
268. Pang W, Tam SC, Zheng YT. Current peptide HIV type-1 fusion inhibitors. *Antivir Chem Chemother* 2009,**20**:1-18.
269. De Clercq E. Antiretroviral drugs. *Curr Opin Pharmacol* 2010,**10**:507-515.
270. Tu X, Das K, Han Q, Bauman JD, Clark AD, Jr., Hou X, *et al.* Structural basis of HIV-1 resistance to AZT by excision. *Nat Struct Mol Biol* 2010,**17**:1202-1209.
271. De Clercq E. Antiviral drug discovery and development: where chemistry meets with biomedicine. *Antiviral Res* 2005,**67**:56-75.
272. D'Cruz OJ, Uckun FM. Dawn of non-nucleoside inhibitor-based anti-HIV microbicides. *J Antimicrob Chemother* 2006,**57**:411-423.
273. Jamjoom GA, Azhar EI, Madani TA, Hindawi SI, Bakhsh HA, Damanhouri GA. Genotype and antiretroviral drug resistance of human immunodeficiency virus-1 in Saudi Arabia. *Saudi Med J* 2010,**31**:987-992.
274. Ren J, Bird LE, Chamberlain PP, Stewart-Jones GB, Stuart DI, Stammers DK. Structure of HIV-2 reverse transcriptase at 2.35-A resolution and the mechanism of resistance to non-nucleoside inhibitors. *Proc Natl Acad Sci U S A* 2002,**99**:14410-14415.
275. de Bethune MP. Non-nucleoside reverse transcriptase inhibitors (NNRTIs), their discovery, development, and use in the treatment of HIV-1 infection: a review of the last 20 years (1989-2009). *Antiviral Res* 2010,**85**:75-90.

276. Johnson VA, Calvez V, Gunthard HF, Paredes R, Pillay D, Shafer RW, *et al.* Update of the drug resistance mutations in HIV-1: March 2013. *Top Antivir Med* 2013,**21**:6-14.
277. Clavel F, Hance AJ. HIV drug resistance. *New England Journal of Medicine* 2004,**350**:1023-1035.
278. Quashie PK, Mesplede T, Wainberg MA. Evolution of HIV integrase resistance mutations. *Curr Opin Infect Dis* 2013,**26**:43-49.
279. Messiaen P, Wensing AM, Fun A, Nijhuis M, Brusselaers N, Vandekerckhove L. Clinical use of HIV integrase inhibitors: a systematic review and meta-analysis. *PLoS One* 2013,**8**:e52562.
280. Kessl JJ, Jena N, Koh Y, Taskent-Sezgin H, Slaughter A, Feng L, *et al.* Multimode, cooperative mechanism of action of allosteric HIV-1 integrase inhibitors. *J Biol Chem* 2012,**287**:16801-16811.
281. Temesgen Z, Talwani R, Rizza SA. Dolutegravir, an HIV integrase inhibitor for the treatment of HIV infection. *Drugs Today (Barc)* 2014,**50**:7-14.
282. Henrich TJ, Kuritzkes DR. HIV-1 entry inhibitors: recent development and clinical use. *Curr Opin Virol* 2013,**3**:51-57.
283. Kuritzkes DR. HIV-1 entry inhibitors: an overview. *Curr Opin HIV AIDS* 2009,**4**:82-87.
284. Berkhout B, Eggink D, Sanders RW. Is there a future for antiviral fusion inhibitors? *Curr Opin Virol* 2012,**2**:50-59.
285. International ASSWGoHIVC, Deeks SG, Autran B, Berkhout B, Benkirane M, Cairns S, *et al.* Towards an HIV cure: a global scientific strategy. *Nat Rev Immunol* 2012,**12**:607-614.
286. Kent SJ, Reece JC, Petravic J, Martyushev A, Kramski M, De Rose R, *et al.* The search for an HIV cure: tackling latent infection. *Lancet Infect Dis* 2013,**13**:614-621.
287. Dolgin E. New, intensive trials planned on heels of Mississippi HIV 'cure'. *Nat Med* 2013,**19**:380-381.
288. Vanham G, Buve A, Florence E, Seguin-Devaux C, Saez-Cirion A. What is the significance of posttreatment control of HIV infection vis-a-vis functional cure? *AIDS* 2014,**28**:603-605.
289. Allers K, Hutter G, Hofmann J, Loddenkemper C, Rieger K, Thiel E, *et al.* Evidence for the cure of HIV infection by CCR5Delta32/Delta32 stem cell transplantation. *Blood* 2011,**117**:2791-2799.
290. Li G, Verheyen J, Rhee SY, Voet A, Vandamme AM, Theys K. Functional conservation of HIV-1 gag: implications for rational drug design. *Retrovirology* 2013,**10**:126.
291. Adamson CS, Sakalian M, Salzwedel K, Freed EO. Polymorphisms in Gag spacer peptide 1 confer varying levels of resistance to the HIV- 1 maturation inhibitor bevirimat. *Retrovirology* 2010,**7**:36.
292. Fun A, Wensing AM, Verheyen J, Nijhuis M. Human Immunodeficiency Virus Gag and protease: partners in resistance. *Retrovirology* 2012,**9**:63.
293. Liu Z, Wang Y, Brunzelle J, Kovari IA, Kovari LC. Nine crystal structures determine the substrate envelope of the MDR HIV-1 protease. *Protein J* 2011,**30**:173-183.
294. Chaudhury S, Gray JJ. Identification of structural mechanisms of HIV-1 protease specificity using computational peptide docking: implications for drug resistance. *Structure* 2009,**17**:1636-1648.
295. Tie Y, Boross PI, Wang YF, Gaddis L, Liu F, Chen X, *et al.* Molecular basis for substrate recognition and drug resistance from 1.1 to 1.6 angstroms resolution crystal structures of HIV-1 protease mutants with substrate analogs. *FEBS J* 2005,**272**:5265-5277.

Chapter 2

Functional conservation of HIV-1 Gag: implications for rational drug design

"Simple can be harder than complex: You have to work hard to get your thinking clean to make it simple."

— Steve Jobs

This chapter is an adapted reprint of my article:
Guangdi Li, Jens Verheyen, Soo-Yon Rhee, Arnout Voet, Anne-Mieke Vandamme, Kristof Theys, Functional conservation of HIV-1 Gag: implications for rational drug design, Retrovirology. 2013 Oct 31;10:126.

I proposed the idea, designed the software and drafted the manuscript. The improvement of the paper was supported with substantial help from Kristof Theys, as well as advices and corrections from other coauthors. I sincerely thank Supinya Piampongsant, Fossie Ferreira, Andrea-Clemencia Pineda-Peña, Ricardo Khouri, Mónica Eusébio, Soraya Maria Menezes and Dan Clements for technical assistance and valuable contributions to the analysis.

2.1 Summary

HIV-1 replication can be successfully blocked by targeting *gag* gene products, offering a promising strategy for new drug classes that complement current HIV-1 treatment options. However, naturally occurring polymorphisms at drug binding sites can severely compromise HIV-1 susceptibility to Gag inhibitors in clinical and experimental studies. Therefore, a comprehensive understanding of Gag natural diversity is needed. We analyzed the degree of functional conservation in 10862 full-length Gag sequences across 8 major HIV-1 subtypes and identified the impact of natural variation on known drug binding positions targeted by more than 20 Gag inhibitors published to date. Complete conservation across all subtypes was detected in 147 (29%) out of 500 Gag positions, with the highest level of conservation observed in capsid protein. Almost half (41%) of the 136 known drug binding positions were completely conserved, but all inhibitors were confronted with naturally occurring polymorphisms in their binding sites, some of which correlated with HIV-1 subtype. Integration of sequence and structural information revealed one drug binding pocket with minimal genetic variability, which is situated at the N-terminal domain of the capsid protein. This first large-scale analysis of full-length HIV-1 Gag provided a detailed mapping of natural diversity across major subtypes and highlighted the considerable variation in current drug binding sites. Our results contribute to the optimization of Gag inhibitors in rational drug design, given that drug binding sites should ideally be conserved across all HIV-1 subtypes.

2.2 Introduction

A curative therapy or preventive vaccine for HIV-1 infected patients remains elusive to date. Standard HIV treatment is confronted with the emergence of viral resistance to existing drug classes, necessitating the development of inhibitors with new mechanisms of action [1]. The Gag polyprotein, essential for HIV-1 morphogenesis, comprises four major domains (matrix, capsid, nucleocapsid, p6) and two small spacer peptides (p1, p2) [2]. Recently, HIV-1 inhibitors that target different stages of virion morphogenesis demonstrated promising antiviral activity, mainly by inhibiting capsid assembly, disrupting nucleocapsid binding with viral RNA/DNA or blocking proteolytic processing of polyproteins during maturation [2-5].

HIV-1 subtype B isolates were predominantly used for the *in vitro* experiments. Non-B subtypes however account for 90% of HIV-1 infections worldwide [6] and amino acid (AA) compositions can differ up to 30% between subtypes [7]. Recently, treatment failure of patients in a phase II clinical study of the maturation inhibitor bevirimat was attributed to natural polymorphisms at drug binding positions, showing up in subtype-specific patterns [8]. Studies that extensively investigate the implications of HIV-1 diversity for Gag-directed drug development are lacking to date. In this large-scale analysis, we examined the distribution of naturally occurring sequence variability in full-length Gag sequences of major HIV-1 subtypes. Moreover, we evaluated the impact of HIV-1 subtypes on the conservation of Gag drug binding positions and multisite binding pockets published to date.

2.3 Materials and Methods

We retrieved 12543 *gag* nucleotide sequences spanning all 1500 base pairs from the HIV Los Alamos database (http://www.hiv.lanl.gov). Sequences were aligned against the HXB2 reference and manually curated using Seaview 4.3 [9]. Hypermutated sequences were detected using the Los Alamos hypermut tool [10]. HIV-1 subtype was determined by the Rega [11] and COMET subtyping tools (http://comet.retrovirology.lu/). Sequence quality was ensured by excluding duplicates and sequences with internal stop-codons, hypermutations, more than 1% ambiguous nucleotides, discordant subtype classification or an identical combination of patient code, sampling year and country. The analysis was restricted to the major subtypes and circulating recombinant forms (CRFs) characterizing the global HIV-1 subtype distribution [6]. For each individual subtype, amino acids that differed from the corresponding consensus AA and with prevalence $\geq$ 0.5% were defined as polymorphisms [12]. PDB data of protein-inhibitor complexes were collected from the RCSB Protein Data Bank [13], summarized in Additional file 1. The AA sequences in each PDB were aligned against the HXB2 reference. Drug binding pockets were defined by protein positions within a minimum Euclidean distance of less than 5Å between atoms of inhibitors and non-hydrogen atoms of residues [14]. Information on known Gag candidate inhibitors and binding positions was retrieved from more than 50 publications, summarized in Additional file 1.

To quantify the degree of positional conservation, a conservation index (CI) was calculated for each position by averaging pairwise scores between all AAs using the BLOSUM62 substitution matrix. Adapted from Karlin and Brocchieri [15], the conservation index (CI) of position x is calculated as: $CI(x) = 1 - \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} [S(x_i, x_j) / \sqrt{S(x_i, x_i) S(x_j, x_j)}]$, where x_i is the amino acid at position x in the ith sequence of the multiple sequence alignment (MSA), N is the number of sequences in the MSA and $S(x_i, x_j)$ is the substitution score of BLOSUM62 between amino acids x_i and x_i. Given that denominators cannot be zero, a linear transformation was applied to $S(x_i, x_j)$ by adding the absolute value of the minimum score $|\min(S)| + 1$. CI measures were scaled between 0 and 1, with a CI value of 0 indicating that AA variation was absent at that position. A highly conserved position was identified if its CI is below 0.01 for each HIV-1 subtype, a cutoff which corresponds approximately to cumulative polymorphism prevalence below 1% (Additional file 2). The Mann–Whitney U test was performed to compare CI distributions. Performance of the CI method is evaluated in Additional file 2 and our Matlab toolbox is available in Additional file 4.

2.4 Results

We analyzed 10862 full-length Gag sequences that fulfilled the quality criteria, encompassing 8 HIV-1 group M subtypes and CRFs: A1 (n = 1648), B (n = 4131), C (n = 2780), D (n = 443), F1 (n = 35), G (n = 49), CRF01_AE (n = 1714) and CRF02_AG (n = 62). Sequences were sampled from 61 countries between 1981 and 2012. **Table S2.1** summarizes more than 50 Gag inhibitors including their binding sites, target protein, mechanism of action, HIV-1 subtypes and PDB data. These candidate inhibitors were either small organic molecules or peptides and primarily targeted the capsid or nucleocapsid proteins. A total of 136 Gag positions were reported as drug binding positions, of which 53 interacted with more than one inhibitor.

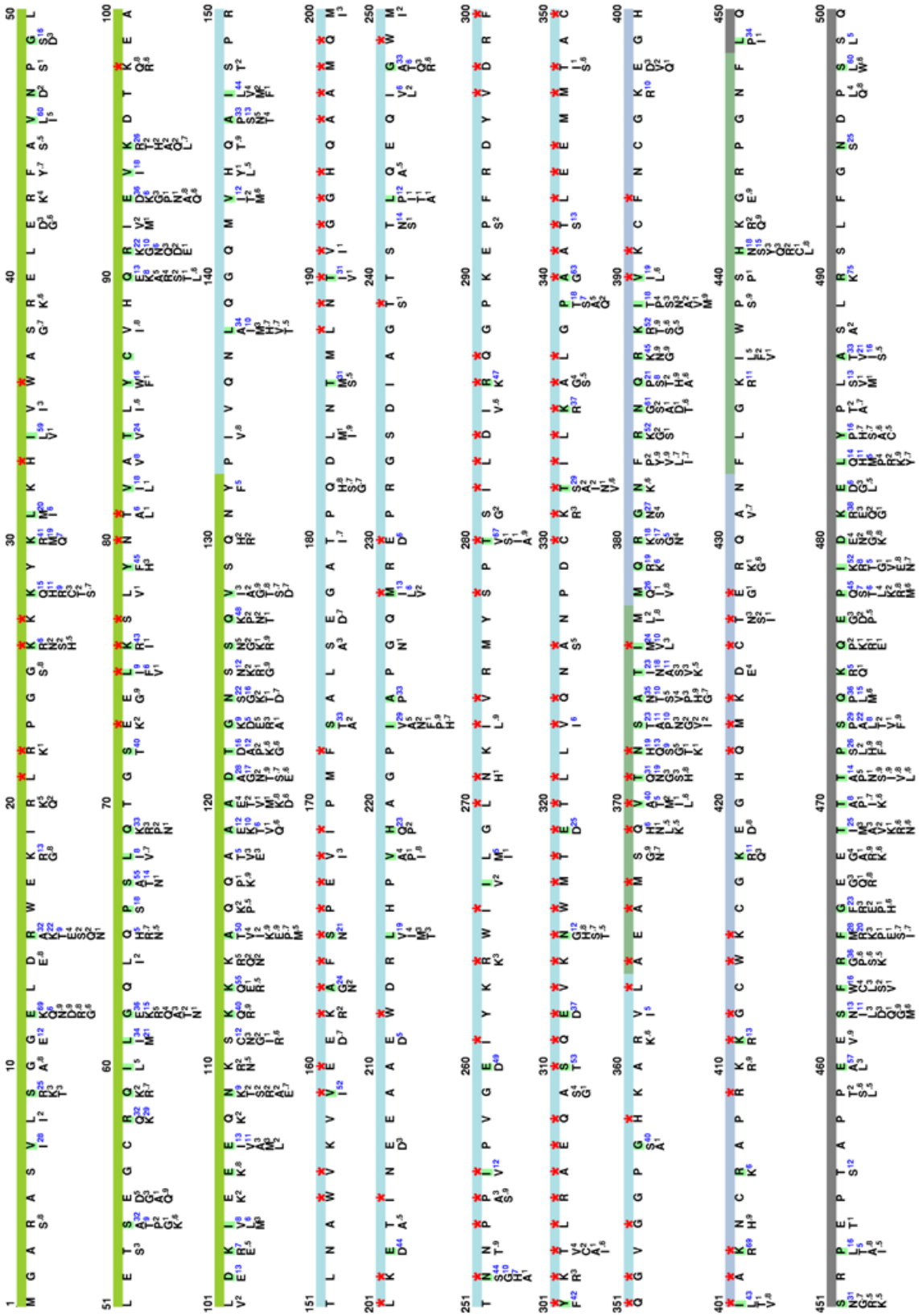


Figure 2.1: Distribution of natural variations at 500 Gag positions of HIV-1 group M (subtypes: A1, B, C, D, F1, G and CRF01_AE, CRF02_AG). The first position of each protein region is labeled with its protein name in a box. Annotated protein regions are indicated as colored bars: light-green for matrix (positions 1–132), light-blue for capsid (133–363), dark-green for p2 (364–377) and p1 (433–448), dark-

blue for nucleocapsid (378–432) and grey for p6 (449–500). HXB2 indices for both full-length Gag and individual proteins are shown on top of the colored bars (e.g. '180|48' indicates the Gag position 180 and the capsid position 48). Known drug binding positions are marked with red stars. Consensus subtype B amino acid for each position is shown directly under the bar, and is highlighted green when the consensus AA differed in one or more subtypes. Natural polymorphisms are shown below the consensus subtype B amino acids; proportions (%) are colored blue for proportion ≥ 5%; orange otherwise. **Figure S2.3** provides the distribution of natural polymorphisms within each individual subtype.

The AA distribution at 500 Gag positions among HIV-1 group M sequences is shown in **Figure 2.1** and subtype-specific distributions are also visualized (**Figure S2.3**). Heterogeneity in consensus sequences was observed at 142 (28.4%) positions across subtypes, while pairwise comparisons of consensus sequences showed an average of 11.6% difference between subtypes. On average, 43.6 ± 2.7% of positions harbored at least one polymorphism relative to its subtype consensus residue (**Table 2.1**). The capsid protein (29.4%) contained the lowest number of polymorphic positions followed by nucleocapsid (42.5%), matrix (59.9%), and p6 (65.6%). Moreover, of 147 conserved positions in Gag, 67.8% were in capsid, 11.2% in nucleocapsid, 10.5% in matrix and 4.6% in p6. Pairwise AA diversity (Additional file 3) of full-length Gag sequences decreased from 17.0 ± 1.6% between subtypes to 9.0 ± 1.0% within subtypes (**Table 2.2**). The mean AA diversity was significantly lower for capsid (5.0 ± 0.8%) than for nucleocapsid (7.9 ± 2.8%), matrix (13.2 ± 2.0%) or p6 (14.7 ± 2.0%) (p-value<0.05) (**Table 2.3**). The CI distributions of full-length Gag characterized three conserved regions located at the nucleocapsid zinc-finger domains, the capsid N-terminal domain (NTD) and C-terminal domain (CTD) (**Figure 2.2**).

Each Gag protein is demonstrated with its total AA and drug binding positions (e.g. Matrix [132/13]: 132 AA positions/13 drug binding positions). Natural polymorphism proportion (%) is indicated for each protein and subtype with respect to the total number of AAs and to the number of drug binding positions (e.g. 57.6/38.5 shows that for subtype B matrix, 57.6% of all 132 positions and 38.5% of 13 drug binding positions are polymorphic). Mean values across Gag domains and across HIV-1 subtypes are indicated in the last row and column respectively.

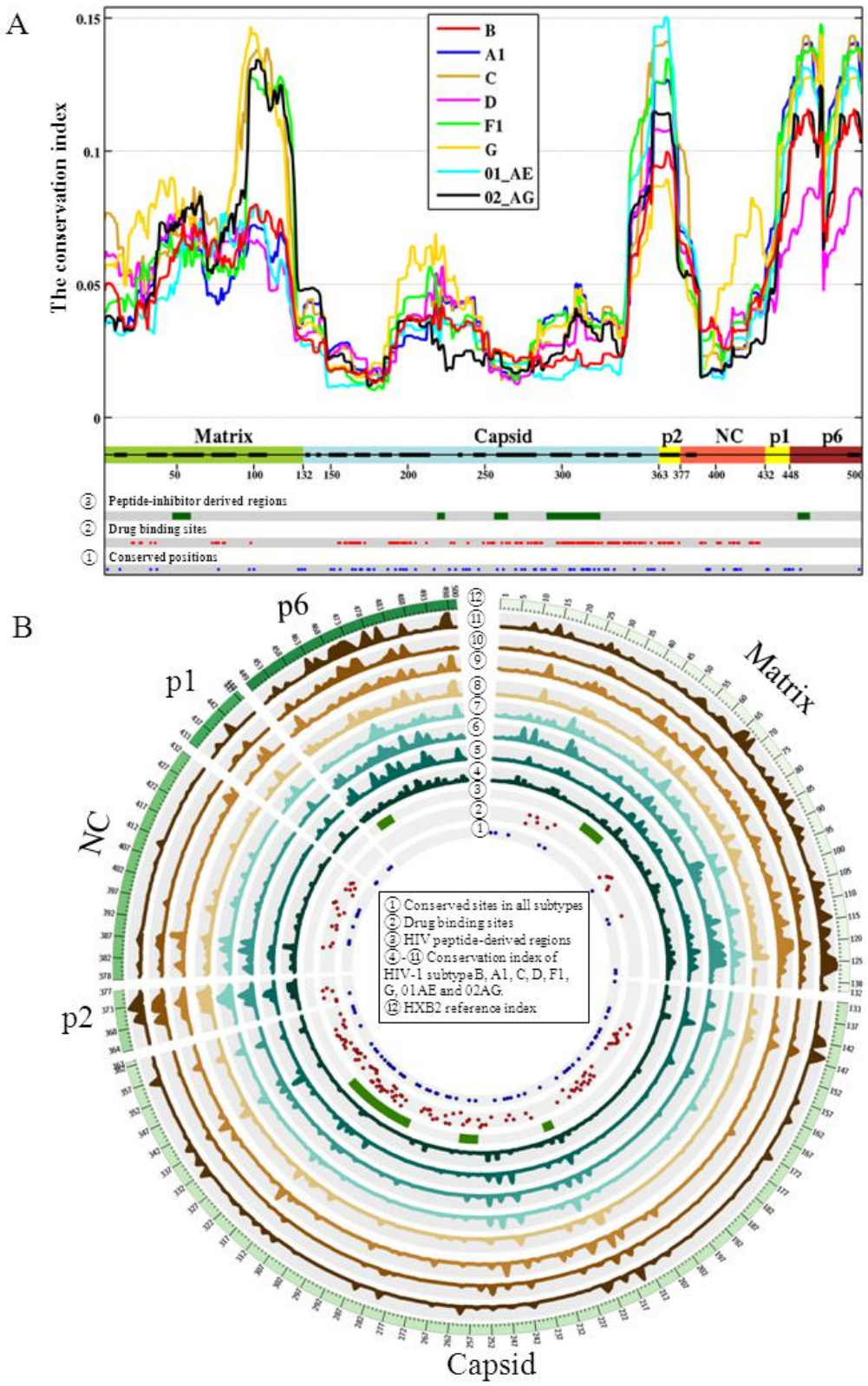


Figure 2.2: Amino acid conservation in the HIV-1 full-length Gag. (A) Sliding window plots of CI values are shown for 8 HIV-1 subtypes (window size: 30AA, also see the plots of exact CI values in **Figure S 2.9**). Secondary structures are indicated for each protein region, with thick lines for helices and thin lines for random-coil structures. Positions conserved in all subtypes are colored blue (layer 1 in a small circle), known drug binding positions are colored red (layer 2) and regions where HIV-1 peptide inhibitors have been derived are colored green (layer 3). Figure S2.9

visualizes the conservation index of 500 Gag positions for 8 HIV-1 subtypes. **(B)** Distributions of CI values at 500 Gag positions across 8 HIV-1 subtypes and CRFs. Visualization software: Circos v0.64 (http://circos.ca/).

Table 2.1: Natural polymorphism proportions in Gag domains and drug binding positions across 8 HIV-1 subtypes and CRFs (%)

	B	A1	C	D	F1	G	01_AE	02_AG	Mean
Matrix[132/13]	57.6/38.5	62.1/46.2	59.1/46.2	64.4/53.8	52.3/46.2	66.7/61.5	61.4/46.2	56.1/30.8	59.9/46.2
Capsid[231/98]	27.3/30.6	34.2/33.7	29.4/29.6	27.7/27.6	31.2/37.8	30.3/28.6	28.1/28.6	27.3/27.6	29.4/30.5
p2[14/8]	71.4/62.5	64.3/62.5	64.3/62.5	64.3/62.5	57.1/62.5	71.4/62.5	64.3/62.5	50.0/50.0	63.4/60.9
NC[55/17]	56.4/58.8	41.8/35.3	38.2/41.2	36.4/29.4	34.5/23.5	54.5/58.8	43.6/35.3	34.5/41.2	42.5/40.4
p1[16/0]	37.5/-	25.0/-	31.2/-	43.8/-	25.0/-	31.2/-	31.2/-	12.5/-	29.7/-
p6[52/0]	76.9/-	69.2/-	69.2/-	55.8/-	65.4/-	61.5/-	69.2/-	57.7/-	65.6/-
Mean	45.2/36.8	46.6/36.8	43.4/34.6	42.8/32.4	41.2/38.2	47.0/37.5	44.0/33.1	39.0/30.9	43.6/35.0

Table 2.2: Inter- and intra-subtype diversity of Gag AA sequences in 8 HIV-1 subtypes and CRFs (%)

	Subtype	B	A1	C	D	F1	G	01_AE	02_AG
Intra-subtype		8.96	8.34	9.89	8.91	9.45	10.90	7.58	8.26
Inter-subtype	B		17.54	18.38	12.70	16.52	18.46	17.22	18.61
	A1			17.65	16.73	16.93	17.27	12.71	14.69
	C				16.67	17.21	18.02	18.22	19.59
	D					16.55	17.56	16.85	18.72
	F1						15.93	16.48	17.68
	G							17.70	18.81
	01_AE								14.92

Subtype-specific AA prevalence at the 136 drug binding positions is shown in **Figure 2.3**. Most positions were located within capsid (72.1%) followed by nucleocapsid (12.5%), matrix (9.6%) and p2 (5.9%). Of these positions, 41.2% were conserved across all subtypes, while 20.6% showed a different consensus AA in one or more

subtypes. On average, 33.8% of drug binding positions harbored at least one polymorphism and 16.3% had at least one polymorphism above 5% prevalence. Non-B subtypes displayed 32 polymorphisms at 20 binding positions that were absent in subtype B. Every inhibitor had at least one polymorphic binding position and 15 inhibitors had more than 50% of drug binding positions showing natural polymorphisms. Among all inhibitors, PF-3450074 [16] targeted the most conserved binding positions at the capsid N-terminal domain, with only one being polymorphic (T107A/S ≤ 6.2%) (**Table S2.2**).

Table 2.3: Pairwise AA diversity of Gag domains in 8 HIV-1 subtypes and CRFs (%)

	Matrix	Capsid	p2	NC	p1	p6	Gag
B	12.36	4.56	20.65	10.31	4.77	15.74	8.84
A1	10.69	5.46	13.97	4.75	5.60	17.17	8.18
C	14.77	5.77	23.79	9.79	4.80	15.83	9.96
D	12.74	4.87	25.62	9.43	8.89	10.17	8.68
F1	12.71	5.37	29.24	9.00	6.71	15.32	9.32
G	17.51	6.15	16.06	10.53	8.42	13.90	10.71
01_AE	11.36	3.44	24.36	4.95	7.74	14.07	7.46
02_AG	13.67	4.91	18.50	3.31	5.39	14.39	8.35
Mean	13.17±2.01	5.04±0.79	21.49±4.83	7.92±2.77	6.39±1.60	14.70±1.99	9.02±1.09

Finally, we analyzed known crystal structures of 9 protein-inhibitor complexes, with 8 inhibitors targeting a total of 75 positions (binding pockets 1–4) in capsid and one targeting 23 positions in nucleocapsid (binding pocket 5) (**Figure 2.4**, **Figure S2.4**). Natural polymorphisms with prevalence ≥ 5% were observed in 28 positions of the binding pockets. Conserved positions were observed in 56% of the capsid binding pockets and 43% of the nucleocapsid binding pocket. Pocket 1 (0.0024) had the lowest average CI values compared to pocket 2 (0.008), 3 (0.0216), 4 (0.0337) or 5 (0.0369).

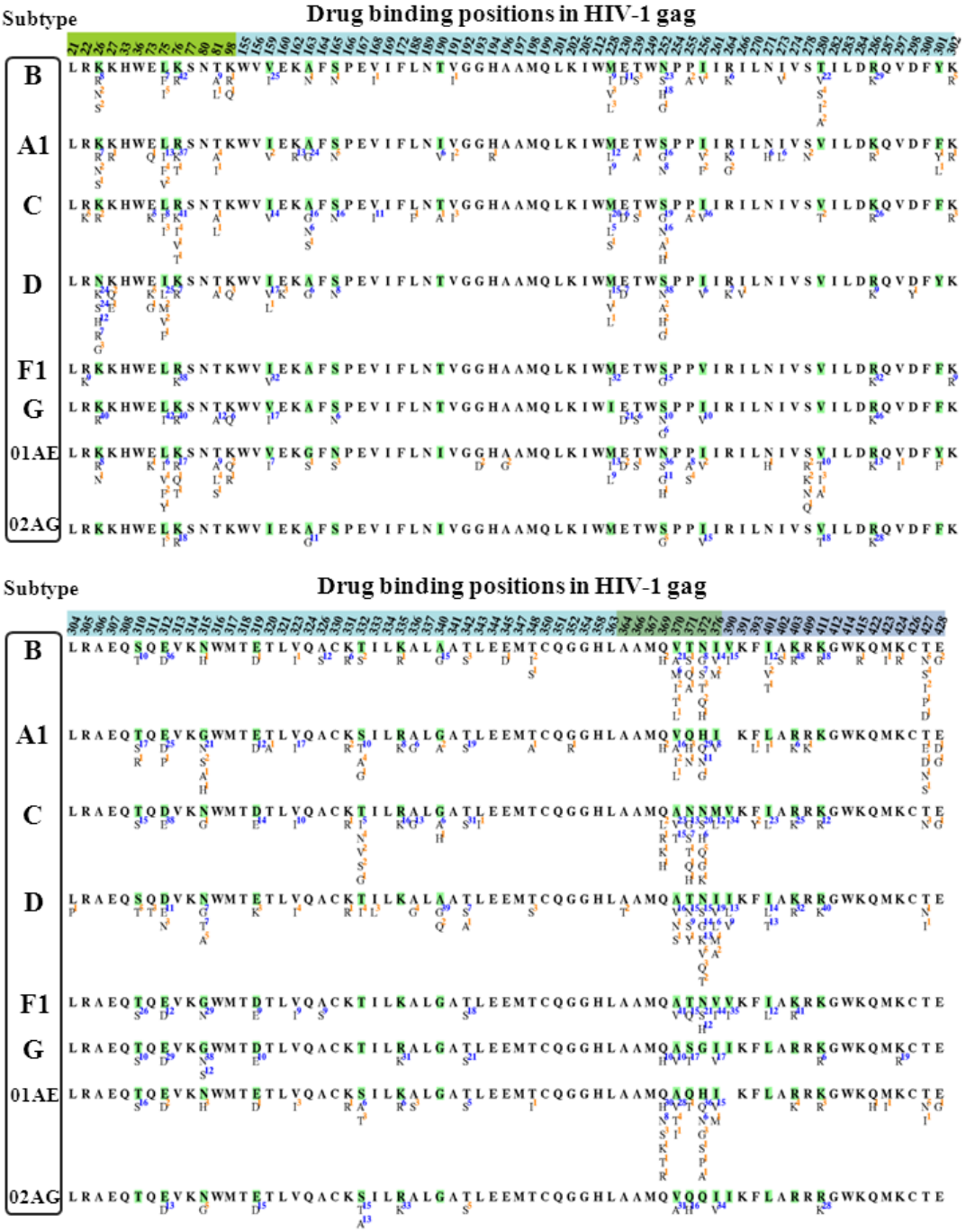


Figure 2.3: Natural polymorphisms at 136 drug binding positions in 8 HIV-1 subtypes and CRFs. For each Gag position, the HXB2 index is shown at the top, followed by the consensus amino acid and natural polymorphisms. Polymorphisms with proportions ≥ 5% are indicated with blue superscripts; orange otherwise.

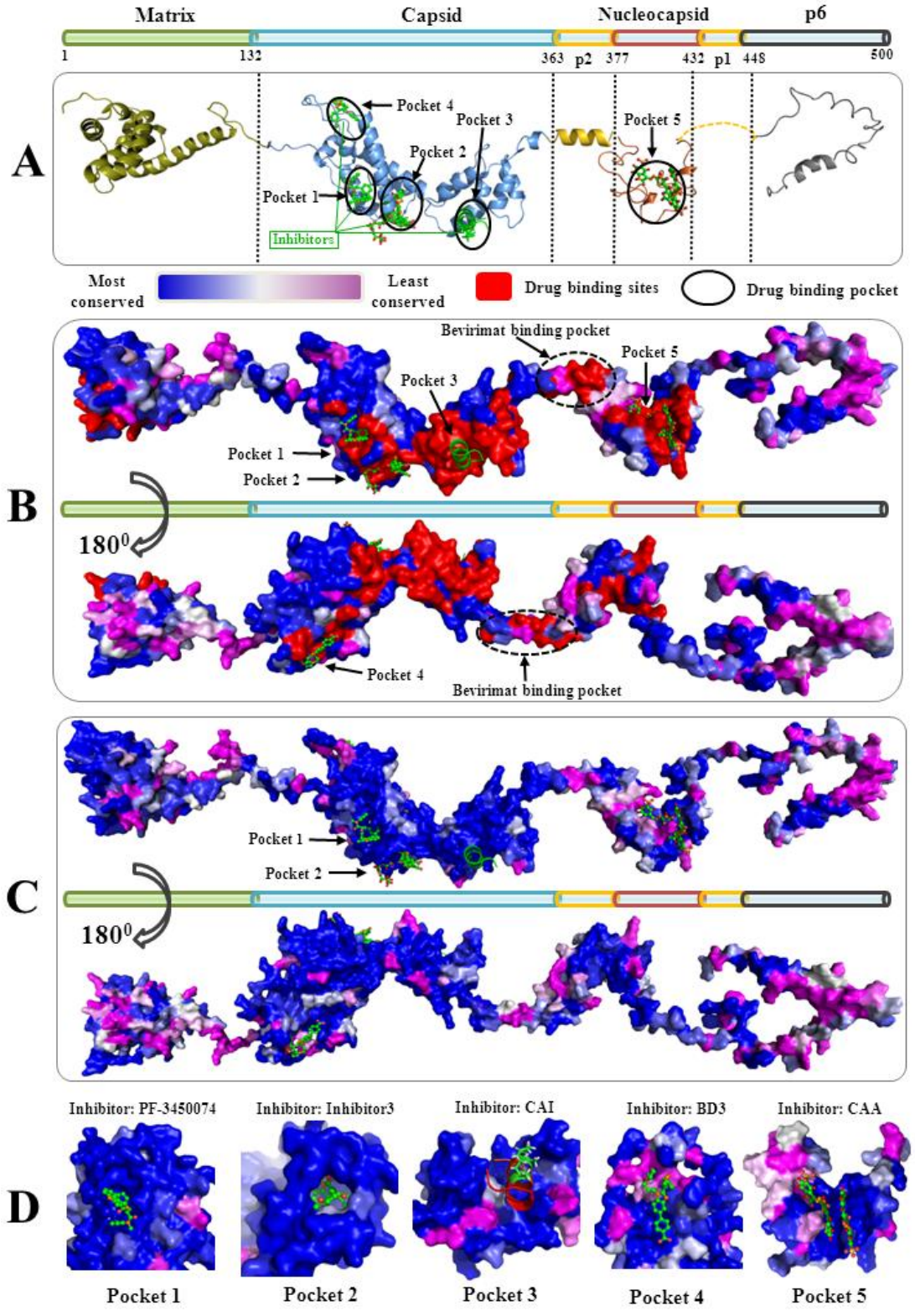


Figure 2.4: Mapping of drug binding positions and binding pockets to HIV-1 Gag protein monomers. The surface spectrum colors indicate the most to the least conserved positions in subtype B from blue CI = 0 to pink CI ≥ 0.1. **(A)** Secondary structures of 4 Gag proteins and 2 spacer peptides, annotated with five drug binding pocket locations. Gag proteins in cartoon representation are colored olive for matrix, blue for capsid, yellow for nucleocapsid, grey for p6, gold for p1 and p2. Bound

inhibitors are represented in green sticks. **(B)** Mapping of drug binding positions to a surface representation of Gag structure, with front and back views. Hypothesized binding positions of bevirimat are also annotated; known drug binding positions are colored red. **(C)** Surface representation of Gag conservation in HIV-1 subtype B (**Figure S2.5** illustrates other subtypes). **(D)** Surface representations of five drug binding pockets in HIV-1 subtype B (**Figure S2.4** shows other subtypes). Inhibitor names are annotated according to publication (**Table S1**). PDB entries of Gag proteins: matrix, 1HIW; capsid, 3NTE; p2, 1U57; nucleocapsid, 2M3Z; p6, 2C55. PDB data of capsid inhibitors: 2BUO, 2L6E, 2XDE, 4E91, 4E92, 2JPR and 4INB, each of which was superimposed to 3H4E using PDBs of 5 drug binding pockets: pocket 1, 2XDE; pocket 2, 4INB; pocket 3, 2BUO; pocket 4, 4E91; pocket 5, 2M3Z. PyMOL V1.5 (http://www.pymol.org/).

2.5 Discussion and conclusions

To our knowledge, our large-scale analysis provided the first detailed mapping of functional conservation of Gag across major HIV-1 subtypes, with implications for the rational design of Gag inhibitors. With more than 50 Gag inhibitors published to date, targeting virion morphogenesis is considered a potential new drug class for HIV-1 treatment [2]. A clinical proof-of-concept was demonstrated in a phase II clinical trial of the maturation inhibitor bevirimat [17], which blocks proteolytic processing at the capsid-p2 cleavage site [18]. Lack of response was observed in 50% of patients and attributed to naturally occurring polymorphisms in the p2 region [8]. A single polymorphism V370A is sufficient for a 40-fold reduction in Bevirimat drug susceptibility [19], with A370 representing the consensus amino acid in several non-B subtypes. Natural diversity was also observed to affect drug effectiveness of other experimental Gag inhibitors [20-22]. Polymorphisms T190I, E230D and I256V, for instance, reduced drug susceptibility to the benzodiazepine and benzimidazole compounds [20]. Moreover, known HIV vaccine candidates containing subtype B Gag gene in HIV-derived vectors did not show sufficient protective efficacies in several large-scale clinical trials [23]. The high diversity of *gag* and *env* genes within and between subtypes can contribute to the challenges of designing a global HIV vaccine neutralizing all HIV-1 subtypes [24]. For the development of HIV vaccine and a potential new drug class targeting virion morphogenesis [2], an assessment of Gag functional conservation and polymorphisms at known drug binding positions is warranted.

We found that 23.4% of drug binding positions in the full-length Gag showed natural polymorphisms in non-B subtypes which could not be detected in subtype B. More importantly, all Gag inhibitors had at least one polymorphic binding position irrespective of subtype. We also found levels of Gag intra- and inter-subtype diversity (9.04% and 17.0%) that exceeded diversity estimates of key viral enzymes (< 7% and < 11%) targeted by standard HIV-1 treatment [12]. However, the most conserved Gag protein capsid has the same level of intra-subtype diversity as integrase (~5%) [12], favoring it as a conserved drug target.

The capsid protein targeted by most candidate inhibitors accounted for 67.7% of conserved Gag positions and contained 72.1% of the 136 binding positions previously reported. Our sequence analysis identified two conserved capsid regions (**Figure 2.2**) located at the interaction interfaces between N-terminal domains (NTD-NTD) as well as between N-terminal and C-terminal domains (NTD-CTD) (). These interaction interfaces, crucial for the assembly and stabilization of pentamer and hexamer lattices [25], provide potential conserved drug targets. To reveal the ideal drug target, we described 4 crystalized drug binding pockets in capsid (**Figure 2.4**, Additional file 3: **Figure S2.4**). Inhibitors that target pockets 1–3 have shown promising antiviral activity against capsid multimerization in different subtype strains by altering NTD-CTD interaction (pockets 1 and 3) or NTD-NTD interaction (pocket 2) [22, 26, 27]. Pocket 4 is less conserved and its polymorphic residues make direct contact with inhibitors, hindering the development of inhibitors that target this pocket [20].

Another potential drug target is the nucleocapsid protein, containing two critical zinc-finger domains for binding with viral RNA genomes [2]. Our conservation analysis mapped the conserved nucleocapsid regions to zinc-finger domains (**Figure 2.2** and **Figure 2.5**) and confirmed previous findings of absolute conservation of CCHC motifs at zinc-coordinating positions [28]. However, we detected considerable variation at other positions, which may alter drug binding and affect antiviral activity. Furthermore, nucleocapsid inhibitors tend to suffer from limited specificity and high toxicity due to the ubiquitous presence of zinc finger domains in many human proteins [4].

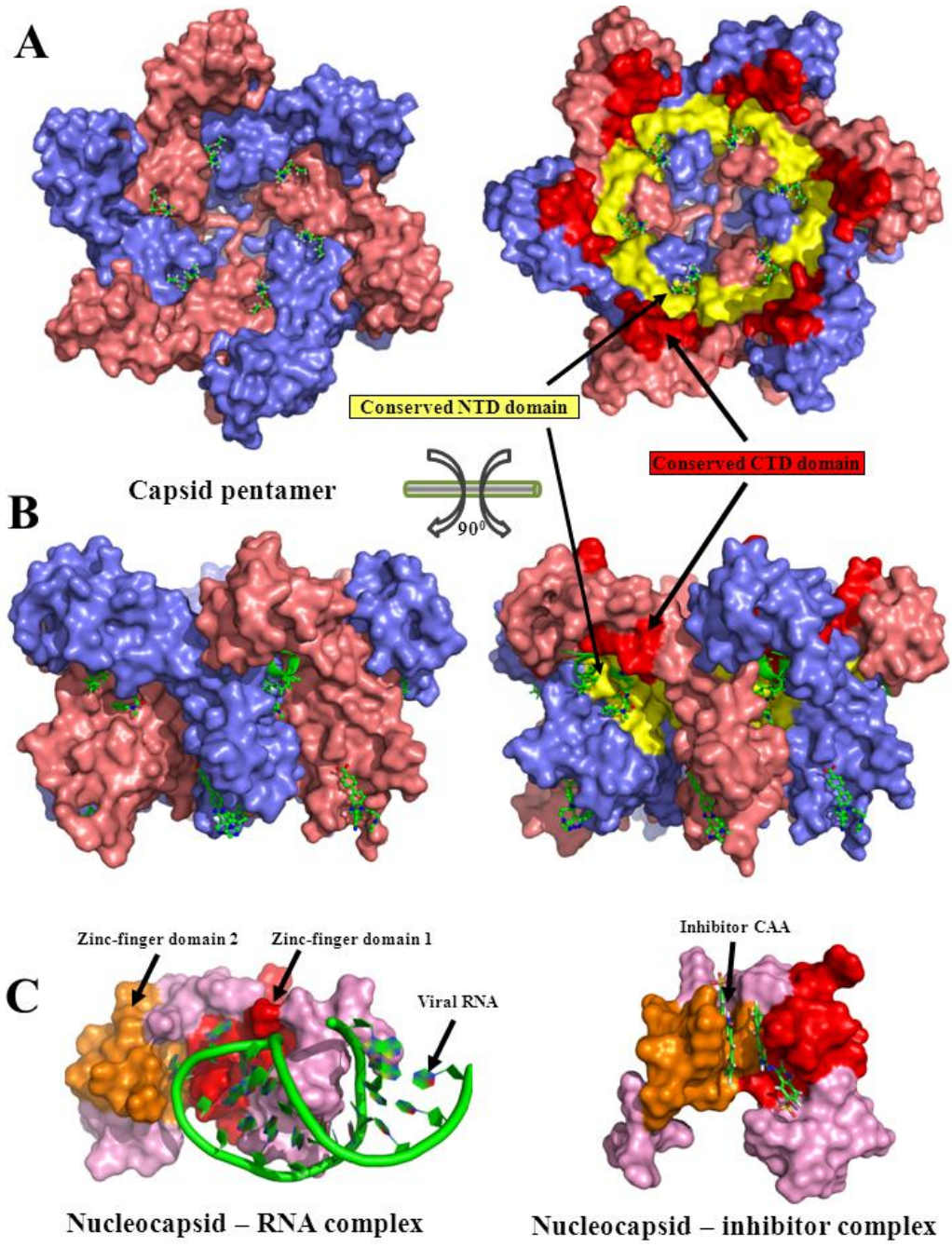


Figure 2.5: Visualization of conserved regions in capsid and nucleocapsid. The capsid hexamer structure (PDB: 3H4E) is shown in top **(A)** and side **(B)** views, with the 6 capsid units (pink, blue), conserved NTD-NTD interaction domains (yellow) and conserved NTD-CTD interaction domains (red). **(C)** The structural complex of nucleocapsid and RNA (left, PDB: 1A1T) and the structural complex of nucleocapsid and inhibitor CAA (right, PDB: 2M3Z). The first zinc-finger domain (nucleocapsid positions: 14–29, Gag positions: 389–404) and the second zinc-finger domain (nucleocapsid positions: 35–50, Gag positions: 410–425) are colored red and orange,

respectively. **Figure S2.7** and **Figure S2.8** provide detailed structures of conserved Gag regions.

Matrix inhibitors with broad spectrum antiviral activities were recently reported, but mutations at drug binding positions significantly reduced their effectiveness [29, 30]. We also observed many natural variants at their drug binding sites (**Table S2.2**), suggesting that further optimization of matrix inhibitors is needed.

Studies that analyzed genetic variability and drug binding site heterogeneity in Gag using large-scale sequence populations are lacking. Previously, small subtype B sequence datasets were used to characterize Gag conservation (n = 125) [31] or positive selective pressure (n = 635) [32]. Polymorphisms at drug binding sites of capsid inhibitor PF-3450074 [16] and conservation of nucleocapsid zinc-finger domains [28] were also reported using fewer than 200 sequences. The only large-scale analysis that we found [33] quantified the drug binding site conservation of a single matrix inhibitor and lacked information on subtype-specific variations. By contrast, we presented here a large-scale and integrative analysis using 10862 full-length Gag sequences, 136 Gag inhibitor drug binding positions and 14 PDB structures. Natural polymorphisms of full-length Gag were detected across 8 major HIV-1 subtypes and a robust estimation of functional conservation was performed using CI analysis, which incorporated biochemical similarities between amino acids (Additional file 3). This sequence analysis predicted three conserved drug targets in Gag (**Figure 2.2**) which were confirmed by existing structural knowledge (**Figure 2.5**).

This study is limited in that it neither addressed how to optimize known Gag inhibitors nor quantified the impact of newly identified polymorphisms on antiviral activities of investigated inhibitors. We collected all available PDBs of Gag-inhibitor structures from the RCSB protein data bank, but more crystallized complexes are needed to reveal novel mechanisms of action. Moreover, the limited number of available Gag sequences for subtypes F1, G and CRF02_AG (n < 100) may have affected the identification of polymorphic positions, but consistent conservation patterns were observed in Gag regardless of HIV-1 subtype (**Figure 2.2**). While we attempted to be as comprehensive as possible, additional inhibitors may have been reported. Conservation of their binding positions can nevertheless be deduced from our full-length Gag analysis. Future studies are also needed to address whether

interactions between Gag and protease can affect Gag drug binding sites, leading to compromised drug activities of Gag inhibitors [34].

In conclusion, our study presented a comprehensive mapping of functional conservation in Gag and strengthened the idea of capsid as a potential target for HIV-1 therapeutics. Increased knowledge on HIV-1 natural diversity in drug binding pockets contributes to rational design of Gag inhibitors and it remains a challenge to design Gag inhibitors with drug binding sites conserved across HIV-1 subtypes.

2.6 Additional file 1: Tables

Table S2.1: **Summary of experimental Gag inhibitors published to date**

	Inhibitor names	Target protein	Binding sites in target protein	Detected effect	HIV strain	PDB	Year	Ref
Peptide inhibitors	Matrix1	Matrix	?	Inhibit assembly and maturation	B		1994	[35]
	Capsid1	Capsid	?	Inhibit assembly and maturation	B		1994	[35]
	CAI	Capsid	D166,Y169,K182,N183,E187,E212,M215	Inhibit capsid CTD interaction	Cell-free	2BUO	2005	[36]
	CAC1/CAC1M,H8	Capsid	I150,R154,Q155,R168,L173,Q179,N184,W185,M186,T189,L191, V192,C199,T201,I202,L203,K204,A205,L212,E213,M215,M216, A218,C219,Q220, V166,L173,E181, N194,D198,K200, G209,A210,T211,V222	Inhibit capsid assembly	B		2011	[37]
	NYAD series	Capsid	V165,F168,Y169,L172,R173,K182,N183,T186,L211,E212,M215[38]	Inhibit capsid CTD interaction	A,B,C,F,G,01AE	2L6E	2011	[26]
	P-1, P-2, P-3, P-4	Capsid	L151,R154, R184, M185, T188, K203(note that only the largest changesof NMR in the presence of peptides were notified)	Inhibit CA polymerization	B		2011	[39]
	CP4	Capsid	T148, I150, L151, D152, V181, A185, T186, Q192, A208, L211, E212, T216	Inhibit CA-hLysRS interaction	Cell-free		2011	[40]
	HAGPIA	CypA	H54,R55,G72,Q63,N102,H126,W121	Inhibit CA-CypA interaction	Cell-free	1AWR	1997	[41]
	Gagp6 (p6:346-354)	TSG101	T58,Y63,R64,Y68,N69,I70,T92, M95,K98,V141,F142,S143	Inhibit TSC101 UEV domain binding with Gag	Cell-free	3OBU	2010	[42]
Small organic molecule	Compound14	Matrix	?	Inhibit PI(4,5)P_2-matrix interaction	A,B,C,D,E,F,G		2012	[33]
	Compound7	Matrix	L21,R22,W36,R76,T81,K98	Inhibit PI(4,5)P_2-matrix interaction	A,B,C,D,E,F,G		2013	[29]
	TD1,TD2,TD3	Matrix	L21,R22,K26,K27,H33,W36,E73, L75,R76,S77	Inhibit MA-RNA interaction	B		2013	[30]
	Bevirimat	CA-SP1	?	Inhibit Gag maturation	B		2003	[43]
	Bevirimat analogs (C-28, C-30)	CA-SP1	H358,L363,A364 A366,Q369,A370 T371	Inhibit Gag maturation	B		2011	[18]
	Vivecon (MPC-9055)	CA-SP1	?	Inhibit Gag maturation	A,B,C,D,E,F,G, group O and N		2009	[44]
	PA1050040	CA-SP1	?	Inhibit Gag maturation	-		2009	[46]
	MPI-461359	CA-SP1	?	Inhibit Gag maturation	-		2010	[47]
	PF-46396	CA-SP1	?	Inhibit Gag maturation	B,C,E[48]		2012	[49]
	Compound(16)	CA-SP1	A366, M367, Q369, V370, N372, I376	Inhibit Gag maturation	B		2012	[50]
	I-XW-053	Capsid	S33,P34,E35,V36,V165,D166,F1	Inhibit NTD interaction	A,B,C,D,E,F,G,		2012	[51]

			68,Y169,K170,T171,L172,R173, A174,E175,Q176,S178,Q179,E180,N183,T186		group O			
	PF-3450074	Capsid	N57,M66,Q67,K70,I73,T107	Inhibit capsid NTD interaction	A,B,C,D,E	2XDE	2010	[16]
	BM4	Capsid	W23,V27A/I,K30R,F32,S33G,T58I,H62	Inhibit capsid NTD interaction	B	4E92	2012	[22]
	BD3	Capsid	W23,V24,V27A/I,F32,V36T,T58I ,V59,G61E,H62	Inhibit capsid NTD interaction	B	4E91	2012	[22]
	Inhibitor4	Capsid	W80,M96,E98,W117,H120,P122, P123,I124,I129,R132	Inhibit capsid NTD interaction	B	4E91	2013	[52]
	BMMP	Capsid	?	Inhibit Gag-Gag interaction	B		2011	[53]
	Compound3,4 ,5	Capsid	?	Inhibit capsid NTD interaction	Cell-free		2013	[54]
	Compound27	Capsid	?	Inhibit capsid NTD interaction	Cell-free		2013	[55]
	CAP-1	Capsid	W23,V27,E28,A31,F32,V59,H62, A64,A65,I141	Inhibit capsid CTD interaction	B	2JPR	2007	[56]
	CAI-compound series	Capsid	V165,Y169,N183,L211,M215	Inhibit capsid CTD interaction	A,B,C,D,F,G, group O		2011	[57]
	Benzodiazepine series 33	Capsid	V27,A31,F32,V59,H62,A65,Y146	Inhibit capsid NTD interaction	Cell-free		2012	[58]
	Inhibitor3	Capsid	W23,V24.V27,E28,K30,A31,F32, S33,P34,V36,I37,F40,K56,V59,G60,G61,H62,A65,M66,K69,I134, K138,N139,I141,V142	Inhibit capsid NTD interaction	B	4INB	2013	[27]
	CAA	NC	V13,K14,F16,I24,A25,K26,R32, G35,W37,K38,Q45,M46,K47	Inhibit NC-RNA/DNA interaction	B	2M3Z	2013	[59]
	WDO-217	NC	?	Zinc ejection	B,HIV-2,SIV		2012	[60]
	Compound6, Compound8	NC	F16,R32,K34,W37,Q45,M46,K47	Inhibit NC(11-55)-RNA interaction	B		2012	[61]
	Compound45	NC	C49,T50,E51	Inhibit NC-oligonucleotide	B		1999	[62]
	SL3ligands	NC	zinc fingers(15-28,36-49)	Inhibit NC-RNA/DNA interaction	B		2012	[63]
	mONs	NC	zinc fingers	Inhibit NC-DNA interaction	Cell-free		2011	[64]
	SAMT	NC	zinc fingers	Inhibit NC-RNA/DNA	B		2010	[65]
	NV038	NC	zinc fingers	Zinc ejection	B,HIV-2,SIV		2010	[66]
	CO7	NC	zinc fingers	Inhibit NC-DNA interaction	Cell-free		2009	[67]
	Thioesters	NC	zinc fingers	Inhibit NC-RNA interaction	B		2004	[68]
	YS1332D	NC	zinc fingers	Inhibit NC-RNA interaction	B		2003	[69]
	DIBA	NC	zinc fingers	Zinc ejection	B,HIV-2,SIV		2001	[70]
	PATEs	NC	zinc fingers	Target zinc fingers	B		2001	[69]
	NSC 624151	NC	zinc fingers	Target zinc fingers	A,B,C,D,F,HIV-2,SIV		1996	[71]
	SRR-SB3	NC	zinc fingers	Target zinc fingers	B,HIV-2,SIV		1996	[72]
	NOBA	NC	zinc fingers	Target zinc fingers	Cell-free		1998	[73]
	Enantiomers	NC	zinc fingers	Inhibit NC-RNA interaction	Cell-free		2003	[74]
	2-Mercaptobenzamide Thioesters	NC	zinc fingers	Target zinc fingers	Cell-free		2005	[75]
	thiolcarbamates (TICAs),	NC	zinc fingers	Target zinc fingers	B		2002	[76]
	Azodicarbonamide (ADA)	NC	zinc fingers	Target zinc fingers	B,HIV-2		2000	[77]

Notations: (a) peptide inhibitors – amino acid sequences which were designed to inhibit HIV replication, (b) small organic molecule – low molecular weight organic compounds that were designed to bind Gag proteins. Sequences of peptide inhibitors are:

(1) Matrix1[matrix sites: 47-59]: NPGLLETSEGCRQ,

(2) Capsid1[capsid sites:124-133]: IPVGEIYKRW,

(3) CAI: ITFEDLLDYYGPK(Bio[1])CL,

(4) CAC1M [capsid sites:175-193] SESAASSVKAWMTETLLVANTSS,
H8[capsid sites:158-176] KEPFRDYVDRFYKTLRAEQ,
(5) NYAD-201[capsid sites:178-192] AQEVKXWMTXTLLVA
(X= (S)-alpha- (2'-pentenyl)alanine),
(6) P-1[capsid sites: 181-192] VKNWMTETLLRQ,
(7) CP4: cyclo(D-Ala-Ile-Fpa-Arg-Tyr-Trp-D-Ala-D-Ala-Glu)-Lys
(8) HAGPIA [capsid sites:87-92] HAGPIA
(9) Gagp6[p6 sites:5-13]: PEPTAPPEE
As for the information of clinical trials, Bevirimat is under Phase IIb[17], Vivecon was under phase IIa (discontinued) [44], PA1050040 was under phase I [45] and Azodicarbonamide (ADA) was under phaseI/II (discontinued) [4].

Table S2.2: Summary of natural variants at drug binding sites of Gag inhibitors

Inhibitors	Ref	Target	HIV-1 subtypes							
			B(n=4131)	A1(n=1648)	C(n=2780)	D(n=443)	F1(n=35)	G(n=49)	01_AE(n=1714)	02_AG(n=62)
CAI	[36]	CA		F169Y1.2,G183A1.2,G183N20.9,G183S2.4,E187D11.7	D187E14.1	N183A4.8,N183G7.0,N183T7.0,E187K2.7	G183N30.3,G183S3.0,D187E8.8,	G183N37.5,G183H2.1,G183S12.5,D187E10.4	Y169F1.3,N183H2.9	N183G4.9,N183H1.6,E187D14.8
CAC1/CAC1M,H8	[37]	CA	R154K29.1,K199R6.0,T216I2.4,T216S1.1,A194S12.3,T200S2.5,T210S2.7	K154R3.2,V191I17.1,K199R1.5,R203K8.1,A204G6.0,S200A3.7,S200G1.3,S200T10.1,T210S18.7	K154R26.3,V191I10.0,K199R1.2,R203K16.2,A204G12.9,T200N4.4,T200I5.0,T200S2.1,T200V2.4,T210S30.9,L211I1.4	R154K9.1,Q179T3.4,V191I4.3,I201L3.2,A204G4.1,T216S3.4,T200I4.1,T210S6.6	R154K33.3,V191I8.8,K199R2.9,K203R5.9,A204G5.9,L205M2.9,Q219R2.9,G220E2.9,A194S8.8,T200I2.9,A209G2.9,T210S18.2,L211I2.9	R154K46.8,Q179T2.1,V191I4.3,R203K31.9,A204G4.2,A204S2.1,T200N2.1,T200I2.1,T210S20.8	R154K12.7,V191I3.0,K199R1.2,K203R5.7,A204S2.7,S200A5.8,S200T3.4,T210S5.3	R154K27.9,V191I3.3,V191T1.6,K199R3.3,R203K33.3,A204G3.3,L205M3.3,T216S1.6,A194S3.3,S200A13.1,S200H1.6,S200I1.6,S200T14.8,T210S5.0
NYAD series	[26]	CA		F169Y1.2,G183A1.2,G183N20.9,G183S2.4	L211I1.4	N183A4.8,N183G7.0,N183T7.0	G183N30.3,G183S3.0,L211I2.9	G183N37.5,G183H2.1,G183S12.5	Y169F1.3,N183H2.9	N183G4.9,N183H1.6
P-1, P-2, P-3, P-4	[39]	CA	R154K29.1	K154R3.2,R203K8.1	K154R26.3,R203K16.2	R154K9.1,	R154K33.3,T188S5.9,K203R5.9	R154K46.8,T188S2.1,R203K31.9	R154K12.7,K203R5.7	R154K27.9,R203K33.3
Bevirimat analogs (C-28, C-30)	[18]	CA	Q6H1.9,V7A20.8,V7I1.7,V7M6.2,V7T1.1,T8S1.3	Q6H1.8,V7A16.5,V7I1.9,V7L1.4,Q8N1.3,Q8H3.1	Q6R1.0,Q6L1.7,A7T15.5,A7V22.7,N8Q1.1,N8G14.1,N8S7.1,N8T1.5	A1T2.3,A7V18.8,T8N15.4,T8S8.8	Q6K5.9,A7I3.0,A7L3.0,A7T6.1,A7V42.4,T8A3.3,T8N3.3,T8Q16.7,T8S6.7	Q6N2.1,Q6H10.4,Q6K2.1,A7I2.1,A7M2.1,A7V10.4,S8N2.1,S8T16.7	Q6N7.5,Q6H30.3,Q6K1.3,Q6S2.7,A7T4.0,A7V28.1	V7A31.1,V7T1.6,Q8H16.4
CP4	[40]	CA	T148A2.1,T148I2.5,T148S3.6,T148V22.3,A208G15.6,T216I2.4,T216S1.1	G208A1.6	V148T1.8,G208A5.5,L211I1.4	A208Q2.0,A208G38.9,T216S3.4	V148T2.9,G208A5.9,L211I2.9	V148T4.2,D152G2.1	V148I2.5,V148T10.2	V148T18.0,G208A1.6,T216S1.6

Compound7	[29]	MA	K76R42.8,T81A9.1,T81L1.0,K98R1.0	R76K36.7,T81A4.5	R22K5.0,R76I3.8,R76K41.5,R76V1.3	K76R7.2,K98Q3.4	R22K8.8,W36R2.9,R76K38.2,T81A3.0,T81L6.1	L21M2.1,K76R39.6,T81A12.5,T81P2.1,K98Q6.4,K98T4.3	K76R17.4,T81A9.2,T81L4.1,K98Q1.6	K76R18.0,T81A3.3
TD1,TD2,TD3	[30]	MA	K26R8.2,K26N2.3,K26S2.1,L75I4.9,L75F7.5,K76R42.8	K26R6.6,K26N1.8,K26S1.1,L75I12.8,L75F4.5,L75V1.8,R76K36.7	R22K5.0,K26R1.7,E73K5.1,L75I2.8,L75F8.5,R76I3.8,R76K41.5,R76V1.3	N26R6.7,N26G3.0,N26H12.0,N26K24.2,N26S24.2,K27Q2.0,E73K4.8,I75L24.9,I75M2.0,I75V2.0,K76R7.2	R22K8.8,K26R2.9,W36R2.9,L75I5.9,L75F2.9,R76K38.2	L21M2.1,K26R39.6,K27N4.2,E73Q4.2,L75I42.6,K76R39.6	K26R7.7,L75I6.1,L75F1.6,L75Y1.2,L75V4.2,K76R17.4	K26R1.6,K26N3.3,K26S3.3,L75I4.9,L75V1.6,K76R18.0
PF-3450074	[16]	CA	T107S2.8			T107S2.9	T107A4.2,T107S6.2	T107S1.2		
I-XW-053	[51]	CA	K170R4.5,T171V2.6,S178T10.4,E180D36.1	S33N4.7,F169Y1.2,T171A4.7,T171C9.8,T171I1.2,T171V7.7,T178S17.4,E180D25.5,E180P1.1,G183A1.2,G183N20.9,G183S2.4	S33N15.9,V36I11.4,K170R3.2,T171A1.2,T171V4.0,T178S15.1,D180E38.4	S33N7.7,T171V8.8,S178T5.0,Q179T3.4,D180N2.7,D180E11.1,N183A4.8,N183G7.0,N183T7.0	S33N2.9,V36I2.9,K170R9.1,T171A9.1,T171C3.0,T171V6.1,T178S26.5,E180D11.8,G183N30.3,G183S3.0	S33N6.2,T171A4.2,T171C2.1,T171V2.1,T178S10.4,Q179T2.1,E180D29.2,G183N37.5,G183H2.1,G183S12.5	N33S2.7,Y169F1.3,T178S16.1,E180D1.6,N183H2.9	E35K1.6,T171A1.6,T171V1.6,T178S3.3,E180D13.1,N183G4.9,N183H1.6
BM4	[22]	CA	V27I25.1,	I27V2.3,K30R13.3,S33N4.7,I58V6.4	I27V13.9,S33N15.9	I27V17.0,S33N7.7	W23M2.9,I27V33.3,K30R2.9,S33N2.9	V27I16.7,K30N2.1,S33N6.2	V27I7.0,N33S2.7	I27V3.3,K30R1.6,I58V1.6
BD3	[22]	CA	V27I25.1,	I27V2.3,I58V6.4,V59I1.9	I27V13.9,V36I11.4,V59I3.0	I27V17.0	W23M2.9,I27V33.3,V36I2.9	V27I16.7	V27I7.0,G61D2.0	I27V3.3,I58V1.6
Inhibitor4	[52]	CA	M96I8.6,M96L3.4,M96V3.5,E98D11.0,N120H17.9,N120S23.2,P123A1.8,I124V3.6,R132K5.8	M96I8.9,M96L12.2,S120N7.5,S120G16.0,I124F1.8,I124V2.4,R132G2.0,R132K5.5	M96I20.2,M96L5.1,M96S1.1,E98D6.0,S120A3.1,S120N16.6,S120G19.4,S120H1.2,P123A2.2,I124V35.7	M96I15.2,E98D7.0,S120A2.5,S120N38.0,S120G1.1,S120H2.3,I124V6.3,R132K7.3	M96I32.4,M96L2.9,E98D2.9,S120A2.9,S120G14.7,P122Q2.9,V124I2.9,I129M2.9,R132G2.9,R132K2.9	I96M2.1,E98D20.8,E98G2.1,W117R2.1,S120N10.6,S120G6.4,P123A2.1,I124V10.4	M96I13.5,M96L8.5,E98D2.1,N120G10.8,N120S35.8,P123A8.4,P123S4.1,I124V1.8	M96I1.6,M96L1.6,E98D1.6,S120N1.6,S120G4.9,I124T1.6,I124V14.8,I129M1.6,
CAP-1	[56]	CA	V27I25.1,	I27V2.3,A31G24.5,V59I1.9,I141L5.6	I27V13.9,A31N5.6,A31G16.2,V59I3.0	I27V17.0,E28K2.9,A31G6.3	W23M2.9,I27V33.3,E28G2.9,A31G2.9	V27I16.7,A31G2.1,A31S2.1	V27I7.0,G31S1.1,A64G1.5	I27V3.3,A31N1.6,A31G11.5
CAI-compound series	[57]	CA		F169Y1.2,G183A1.2,G183N20.9,G183S2.4	L211I1.4	N183A4.8,N183G7.0,N183T7.0	G183N30.3,G183S3.0,L211I2.9	G183N37.5,G183H2.1,G183S12.5	Y169F1.3,N183H2.9	N183G4.9,N183H1.6
Benzodiazepine series 33	[58]	CA	V27I25.1	I27V2.3,A31G24.5,V59I1.9,S146N1.9	I27V13.9,A31N5.6,A31G16.2,V59I3.0	I27V17.0,A31G6.3	I27V33.3,A31G2.9	V27I16.7,A31G2.1,A31S2.1,S146C4.2	V27I7.0,G31S1.1,S146R1.9,S146K1.6	I27V3.3,A31N1.6,A31G11.5

Inhibitor3	[27]	CA		K30R13.3,A31G24.5,S33N4.7,V59I1.9,N139H5.6,I141L5.6	A31N5.6,A31G16.2,S33N15.9,V36I11.4,V59I3.0	E28K2.9,A31G6.3,S33N7.7,I134V1.4	W23M2.9,E28G2.9,K30R2.9,A31G2.9,S33N2.9,V36I2.9	K30N2.1,A31G2.1,A31S2.1,S33N6.2,N139H2.1	G31S1.1,N33S2.7,G61D2.0,N139H1.2	K30R1.6,A31N1.6,A31G11.5
CAA	[59]	NC	V13I15.3,I24L11.6,I24T1.5,I24V1.8,K26R48.5	R26K6.0	V13I33.7,F16Y1.7,I24L22.7,R26K25.0	I13L13.2,I13V9.1,I24L14.5,I24T13.3,K26R31.7	V13I35.3,I24L11.8,I24V2.9,K26R42.4	K14R2.1,F16Y2.1,L24I2.1,R26K2.1,Q45L2.1,M46I2.1,K47R18.8	R26K3.6	I13L1.6,I13V1.6,L24V1.6,R26K1.6,Q45L1.6
Compound6,Compound8	[61]	NC	K34R17.9	F16Y1.7,K34R11.9	R34K40.1	K34R2.9	F16Y2.1,K34R6.2,Q45L2.1,M46I2.1,K47R18.8	K34R2.9	R34K29.3,Q45L1.6	
Compound45	[62]	NC	T50N4.7,T50I2.2,T50S4.0,E51G1.6	T50D1.3,T50E1.4,E51D1.2	T50N3.1		T50N2.9,T50E2.9,T50P2.9	T50M2.1,T50S2.1,E51D2.1,E51Q2.1	T50N4.6	T50N3.3,E51V1.6
Bevirimat	[43]	CA-p2	Q6H1.9,V7A20.8,V7I1.7,V7M6.2,V7T1.1,N9Q1.6,N9G8.4,N9S6.7,N9T2.6,I13M2.3,I13V14.1	Q6H1.8,V7A16.5,V7I1.9,V7L1.4,H9N11.5,H9Q29.5,H9G1.1,I13V7.8	Q6R1.0,Q6L1.7,A7T15.5,A7V22.7,N9Q4.9,N9H6.0,N9S20.7,M13L12.2	A7V18.8,N9Q3.4,N9G13.6,N9K12.7,N9S14.7,N9T2.3,N9V4.8,I13A1.6,I13L6.3,I13M3.6,I13V18.8	Q6K5.9,A7I3.0,A7L3.0,A7T6.1,A7V42.4,N9G3.0,N9H12.1,N9K6.1,N9S21.2,V13I44.1,V13L2.9	Q6N2.1,Q6H10.4,Q6K2.1,A7I2.1,A7M2.1,A7V10.4,G9N4.3,G9D2.1,G9S2.1,G9T2.1,I13A2.2,I13M2.2,I13V17.8	Q6N7.5,Q6H30.3,Q6K1.3,Q6S2.7,A7T4.0,A7V28.1,H9N5.6,H9Q35.9,H9G2.6,H9P1.3,H9S1.5,I13M1.3,I13V15.4	V7A31.1,V7T1.6,Q9N3.3,I13V34.4

Inhibitor names, references and targets are indicated in the first three columns. The proportions of natural variations for subtype B, A1, C, D, F1, G, 01AE and 02AG were summarized from the 4^{th} to 11^{th} columns. Natural polymorphisms for each inhibitor are summarized for each subtype and they are annotated in the form of "wildtype+position+mutation+proportion". For instance, "R154K29.1" indicates the most prevalent amino acid R at position 154 switches to amino acid K, with the proportion 29.1% in sequence dataset. Note that the most prevalent amino acid is defined as wildtype in our analysis.

2.7 Additional file 2: Notes

Mathematical model of conservation index

We analyzed the degree of positional conservation in the multiple sequence alignment (MSA), taking into account of stereochemical variability between amino acids. Adapted from the conservation analysis in Karlin and Brocchieri [15, 78], a

conservation index (CI) was calculated for each position by averaging pairwise dissimilarity scores between all AAs using BLOSUM62 matrix [79].

Amino acid substitution matrices (e.g. BLOSUM62) are designed for estimating the occurrence of each possible pairwise substitution over evolutionary time. While the genetic code allows the translation of similar codons into the same synonymous or similar AAs, mutating one AA to another AA with substantially different biochemical properties can affect protein folding or activity [80]. In a substitution matrix, the nondiagonal pairwise scores how likely an AA is to be substituted by another in a homologous protein and the diagonal scores indicate how likely one AA is to be substituted at all [81]. For instance, a negatively charged residue like aspartic acid D is more likely to be replaced by the other negatively charged residue glutamic acid E, than it is to be mutated into positively charged histidine H. In BLOSUM62 matrix, D to E is scored 2, while D to H is -1. Adapted from Karlin and Brocchieri [15, 78], conservation index (CI) of position x is calculated as:

$$CI(x) = 1 - \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{S(x_i, x_j)}{\sqrt{S(x_i, x_i) S(x_j, x_j)}}$$

Where x_i is the amino acid form at the position x of the ith sequence in the MSA, N is the number of sequences in MSA, $S(x_i, x_j)$ is the similarity score between amino acid form x_i and x_j. Suggested in Karlin and Brocchieri [15, 78], we adapted the similarity matrix BLOSUM62 to provide the similarity scores for $S(x_i, x_j)$. Since denominators should not be zero, the values of BLOSUM62 M are linearly transformed into positive by adding the absolute value of minimum score $|\min(M)|+1$. In our analysis, the conservation index of positions with less than 20% gaps is calculated, and the amino acid comparisons were restricted to 20 amino acids (e.g. ARNDCQEGHILKMFPSTWYV). Note that if no natural variations exist at conserved position x, then $CI(x) = 0$ otherwise, $0 < CI(x) < 1$. Given BLOSUM62 as the similarity matrix for $S(x_i, x_j)$, it can be shown that $0 \le CI(x) \le 0.9278$. Besides, the relationship between conservation index and pairwise diversity can be described by the Proposition 1, which explains that conservation index is equal to or less than pairwise diversity. Note that pairwise diversity is defined as:

$1-\frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\delta(x_i=x_j)$, where δ denotes the Kronecker symbol, $\delta(x_i=x_j)$ equals 1 if x_i is identical to x_j; otherwise 0.

Proposition 1. Suppose x is a position in MSA and x_i is a polymorphism at x, $P(x_i)$ is the prevalence of x_i, let $CI(x)$ and $Diversity(x)$ denote conservation index and pairwise diversity, respectively, then:

$$CI(x) \le Diversity(x)$$

Proof: Assume that an amino acid similarity matrix S (e.g. BLOSUM62) satisfies $S(x_i,x_i) \ge S(x_i,x_j)$. We have:

$$1-\delta(x_i=x_j) \ge 1-\frac{S(x_i,x_j)}{\sqrt{S(x_i,x_i)S(x_j,x_j)}}$$

It can then be concluded that:

$$\begin{aligned} CI(x) &= 1-\frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\frac{S(x_i,x_j)}{\sqrt{S(x_i,x_i)S(x_j,x_j)}} \\ &= \frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}[1-\frac{S(x_i,x_j)}{\sqrt{S(x_i,x_i)S(x_j,x_j)}}] \\ &\le \frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}(1-\delta(x_i=x_j)) \\ &= Diversity(x) \end{aligned}$$

The above section explained relationship between CI(x) boundary and pairwise diversity; we now discuss the properties of the relationship between CI(x) and accumulative polymorphism prevalence. Let $C(x)$ be the cumulative polymorphism prevalence at position x, a trivial observation can be found as: $CI(x) \le Diversity(x) \le \min[1, \frac{2N}{N-1}(C(x)+C^2(x))]$. This is derived from the following equations that $Diversity(x) = \frac{2N}{N-1}\sum_{a}\sum_{b\neq a}P(x=a)P(x=b)$ and $C(x) = \sum_{P(x=a)\le 0.5}P(x=a)$, where a and b are two AA forms at position x in MSA.

Theoretical results did not yield a precise value for the upper boundary of CI(X) using C(X). We therefore used our HIV-1 Gag datasets to show the relationship between

$CI(x)$ and cumulative polymorphism prevalence regarding the identification of conserved positions. Given the cutoff 0.01 for both $CI(x)$ and cumulative polymorphism prevalence, we compared the results from both measurements. Suppose S_1 is the set of conserved positions given the cutoff of $CI(x)$, S_2 is the set of positions with cumulative polymorphism prevalence less than 0.01. We found that only 5 out of 147 positions in S_1 were different from S_2, and 6 out of 149 positions in S_2 were different from S_1. The two measurements reach up to 95.9% (6/149) common predictions. In other words, using $CI(x)$ tests to identify conserved sites at cutoff 0.01 can approximately guarantee cumulative polymorphism prevalence below 0.01.

Herein, we provide an adapted example from Valdar [81] to compare conservation index with other state-of-the-art conservation methods (i.e. Shannon entropy, Jensen-Shannon diversity, relative entropy, property relative entropy, sum of pairs [82]). We used our Matlab package to calculate Shannon entropy and the python software from Capra and Singh [82] to calculate the other measurements (default settings).

Table S2.3: Comparison of conservation methods given a simple sequence example

Example	Pos1	Pos2	Pos3	Pos4	Pos5	Pos6	Pos7	Pos8	Pos9
Seq1	E	D	D	D	D	D	I	P	D
Seq2	E	D	D	D	D	D	I	P	V
Seq3	E	D	D	D	D	D	I	P	Y
Seq4	E	D	D	D	D	D	I	P	A
Seq5	E	D	D	D	D	D	L	W	T
Seq6	E	D	D	D	E	E	L	W	K
Seq7	E	D	D	D	E	E	L	W	P
Seq8	E	D	D	D	E	E	L	W	C
Seq9	E	D	D	D	E	F	V	S	R
Seq10	E	D	E	F	F	F	V	S	H
Methods									
Conservation index	0	0	0.0665	0.1636	0.3107	0.4006	0.1580	0.5874	0.6730
Shannon Entropy	0	0	0.1412	0.1412	0.4097	0.4472	0.4581	0.4581	1
Property entropy	0	0	0.0418	0.1253	0.1896	0.1703	0.1998	0.4889	0.6355
Jensen-Shannon	0.8367	0.8299	0.8007	0.7621	0.7102	0.6567	0.6507	0.6075	0.5497
Relative Entropy	0.9447	0.9481	0.9048	0.8257	0.7117	0.6143	0.6238	0.6070	0.5363
Property relative entropy	3.1713	3.1713	3.0608	2.8399	2.6729	2.4370	2.2060	1.7668	1.1429
Sum of pairs	5.0000	5.5500	5.1500	3.9722	2.5444	1.9166	1.7277	1.4666	-1.4888

Given the above example with 10 sequences (Seq1 to Seq10), the following order ranks the positions (Pos1 to Pos9) from the most conserved to the least conserved: Pos1 = Pos2 > Pos3 > Pos4 > Pos5 > Pos6, and Pos7 > Pos8 > Pos9. The most conserved positions are Pos1 and Pos2 where there is no mutation. AA change from D to E is more tolerable than from D to F, thus Pos3 is more conserved than Pos4. Pos4, with fewer mutations, is more conserved than Pos5. Pos7 which possesses all hydrophobic I, L and V are more conserved than Pos8 containing P, W, S from different AA groups (aromatic side group, hydrophobic group, polar uncharged side group). Pos9 is the most variable position with all different AAs.

We found that CI was a robust estimation of the conserved sites for three reasons: (1) positions with no natural variations in the MSA have equal CIs. This is not the case with Jensen-Shannon diversity score, for instance. (2) Positions with higher natural variations have higher CIs. This is not the case with property entropy, for instance. (3) The biochemical similarities between amino acids are taken into account. This is not the case with Shannon entropy where all amino acids are treated equally. Regarding the difference between state-of-the art methods, it has been described extensively in [81] [82]. Given 4130 full-length HIV-1 subtype B Gag sequences, **Figure S2.1** demonstrates the distribution of conservation scores in HIV-1 subtype B Gag using conservation index, Shannon entropy and relative entropy. **Figure S2.2** demonstrates the comparison of Shannon entropy and conservation index using full-length protease sequences sampled from 723 HIV-1 subtype B patients, downloaded from HIV Los Alamos Database.

Figure S2.1 and **Figure S2.2** demonstrate that the three methods show similar patterns in full-length Gag conservation analysis, indicating that conservation index may characterize AA conservation and yield similar patterns to entropy measurements. Note that positional conservation methods based on substitution matrices were criticized for not accounting for gaps [81], gaps were treated as missing data in our analysis and only positions with less than 20% gaps were analyzed. Regarding the performance, it is possible that other state-of-the-art methods provide equally ideal estimations of positional conservation by taking into account stereochemical sensitivity, reviewed in [81]. Taken together, our data show that conservation index

provides sufficient statistical power to quantify positional conservation using the BLOSUM substitution matrix.

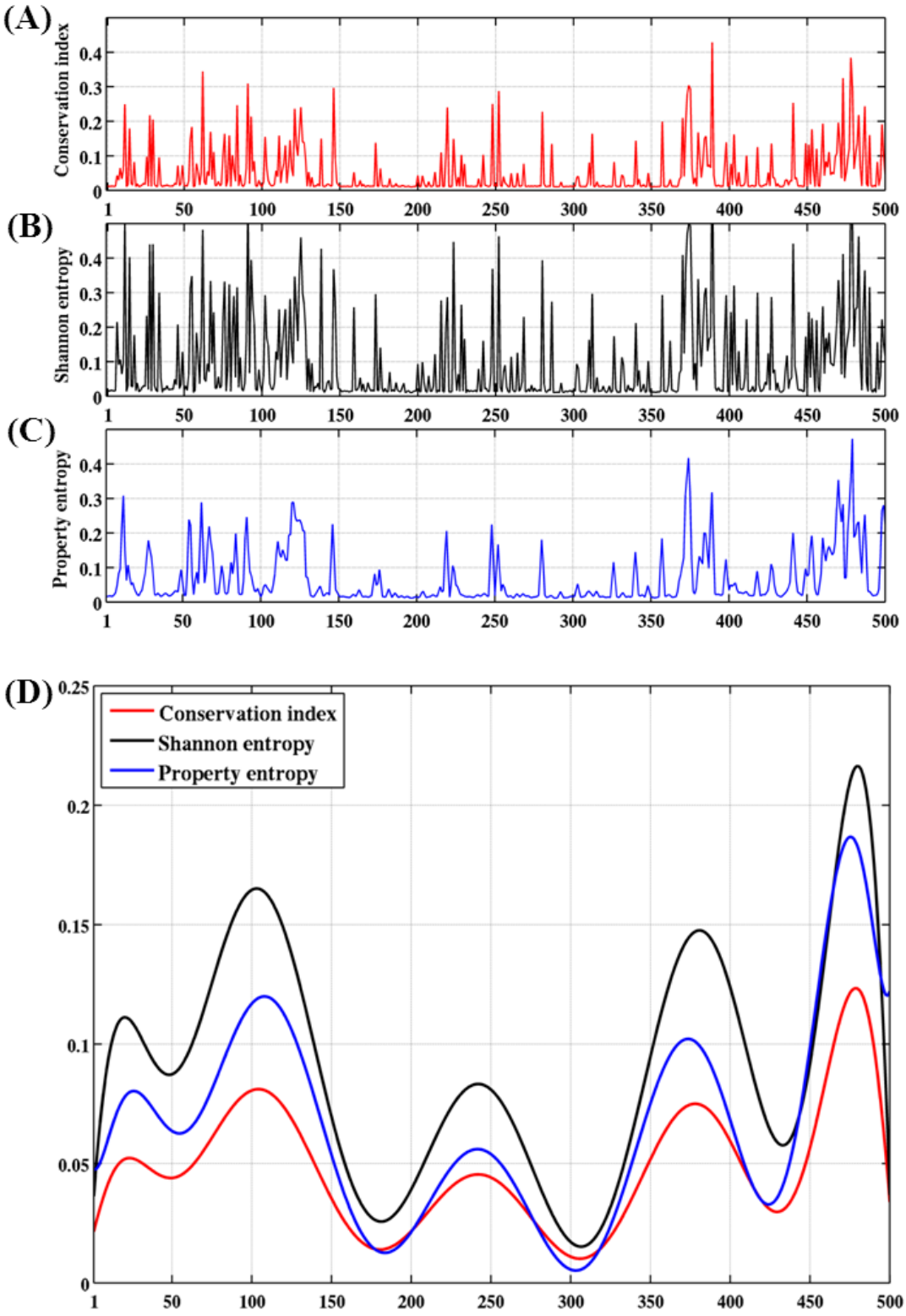


Figure S2.1: **Amino acid conservation in HIV-1 full-length Gag analyzed by conservation index, Shannon entropy and relative entropy.** The plots of the exact

conservation values (A-C) and fitted polynomial curves (D) are provided.

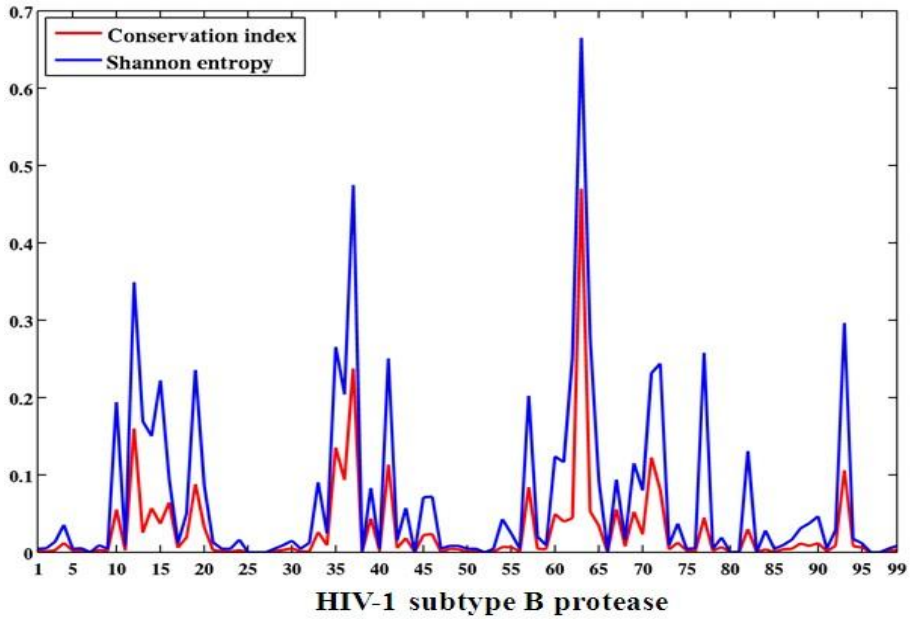


Figure S2.2: HIV-1 protease conservation analyzed by conservation index and Shannon entropy.

Inter- and intra-subtype diversity

The amino acid inter- and intra-subtype diversity was calculated by pairwise amino acid comparisons [12]. Herein we describe the mathematical models. Suppose D is a multiple sequence alignment containing N amino acid sequences, L is the number of positions in D. Intra-subtype diversity $Diversity^{Intra}(D)$ for dataset D is calculated as:

$$Diversity^{Intra}(D) = 1 - \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{1}{L} \sum_{t=1}^{L} \delta(D_i^t = D_j^t)$$

Where D_i^t is the t[th] amino acid form of the sequence i in dataset D, δ denotes the Kronecker symbol, $\delta(D_i^t = D_j^t)$ equals 1 if $D_i^t = D_j^t$ is true; otherwise 0. Similarly, we can calculate the inter-subtype diversity between two sequence datasets. Suppose $D1$ and $D2$ are the multiple sequence alignments from two subtypes (e.g. subtype B and subtype C). Both have the number of sequences, N and M, respectively. The inter-subtype diversity between two subtypes $Diversity^{Inter}(D1, D2)$ is defined as:

$$Diversity^{Inter}(D1, D2) = 1 - \frac{1}{N \times M} \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{1}{L} \sum_{t=1}^{L} \delta(D1_i^t = D2_j^t)$$

In our analysis, we calculated the pairwise diversity at positions on sequences with less than 20% gaps and gaps were treated as missing data. To solve the heavy computation of large sequence datasets (1000 sequences lead to half a million pairwise calculations), our Matlab toolbox implemented parallel computation with optimized memory strategy (http://www.mathworks.com/matlabcentral/fileexchange/47155-amino-acid-conservation-zip).

2.8 Additional file 3: Figures

Figure S2.3: The distribution of natural polymorphisms at 500 Gag positions on 8 HIV-1 subtypes. HXB2 indices at each individual protein are shown in the first column, followed by HXB2 indices in the full-length Gag. Drug binding positions are marked with red stars. The colors in the second column distinguish the Gag domains: light green for matrix (position: 1-132), light blue for capsid (133-363), dark green for p2 (364-377) and p1 (433-448), dark blue for nucleocapsid (378-432) and gray for p6 (449-500). The remaining columns list the consensus amino acid for each subtype followed its natural variation(s) and the corresponding proportion(s) in blue (prevalence above 5%) and orange (prevalence at or below 5%).

Figure S3. Natural variations of gag positions in 8 major HIV-1 subtypes and CRFs

Figure S2.3 (continue)

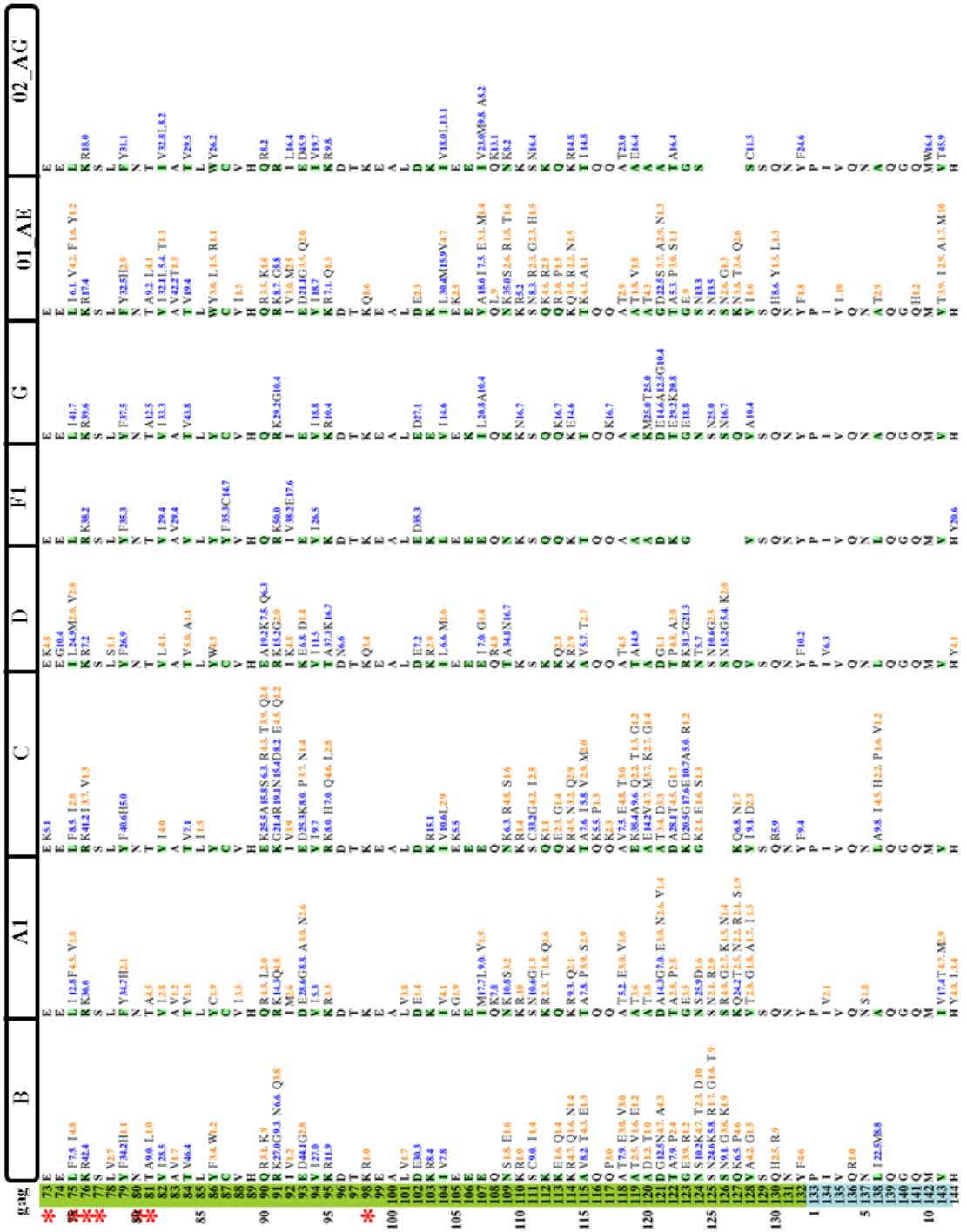


Figure S2.3 (continue)

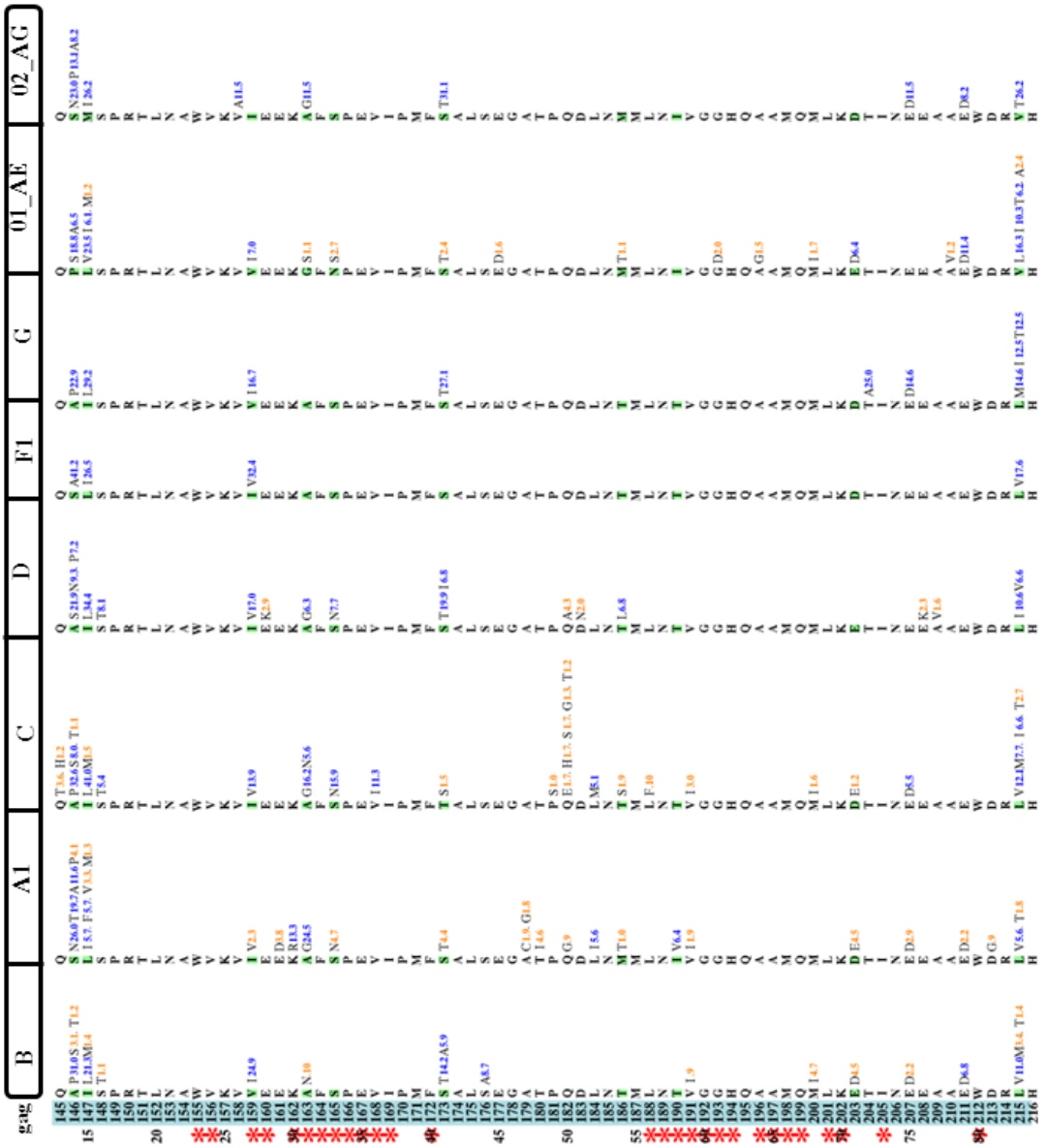


Figure S2.3 (continue)

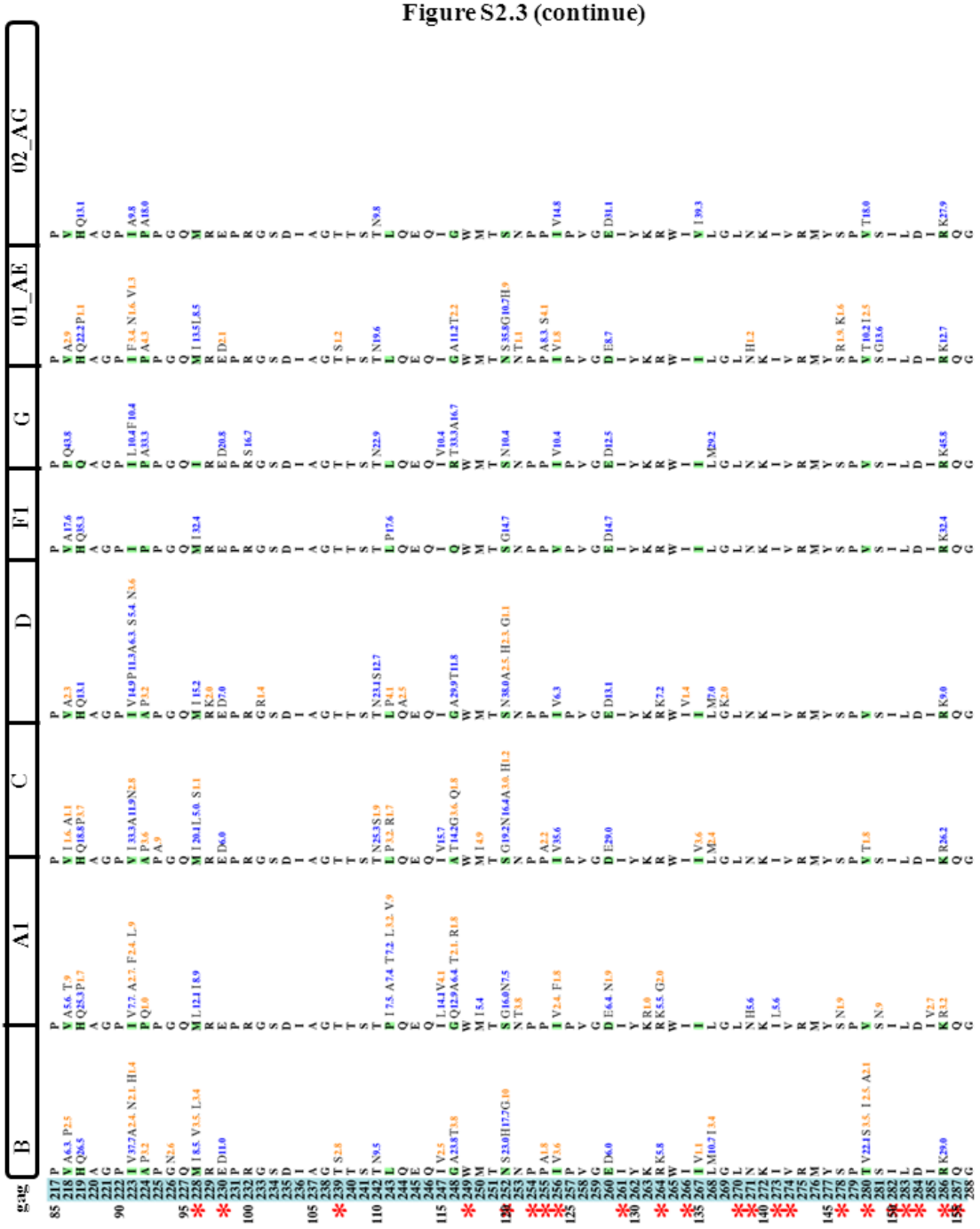


Figure S2.3 (continue)

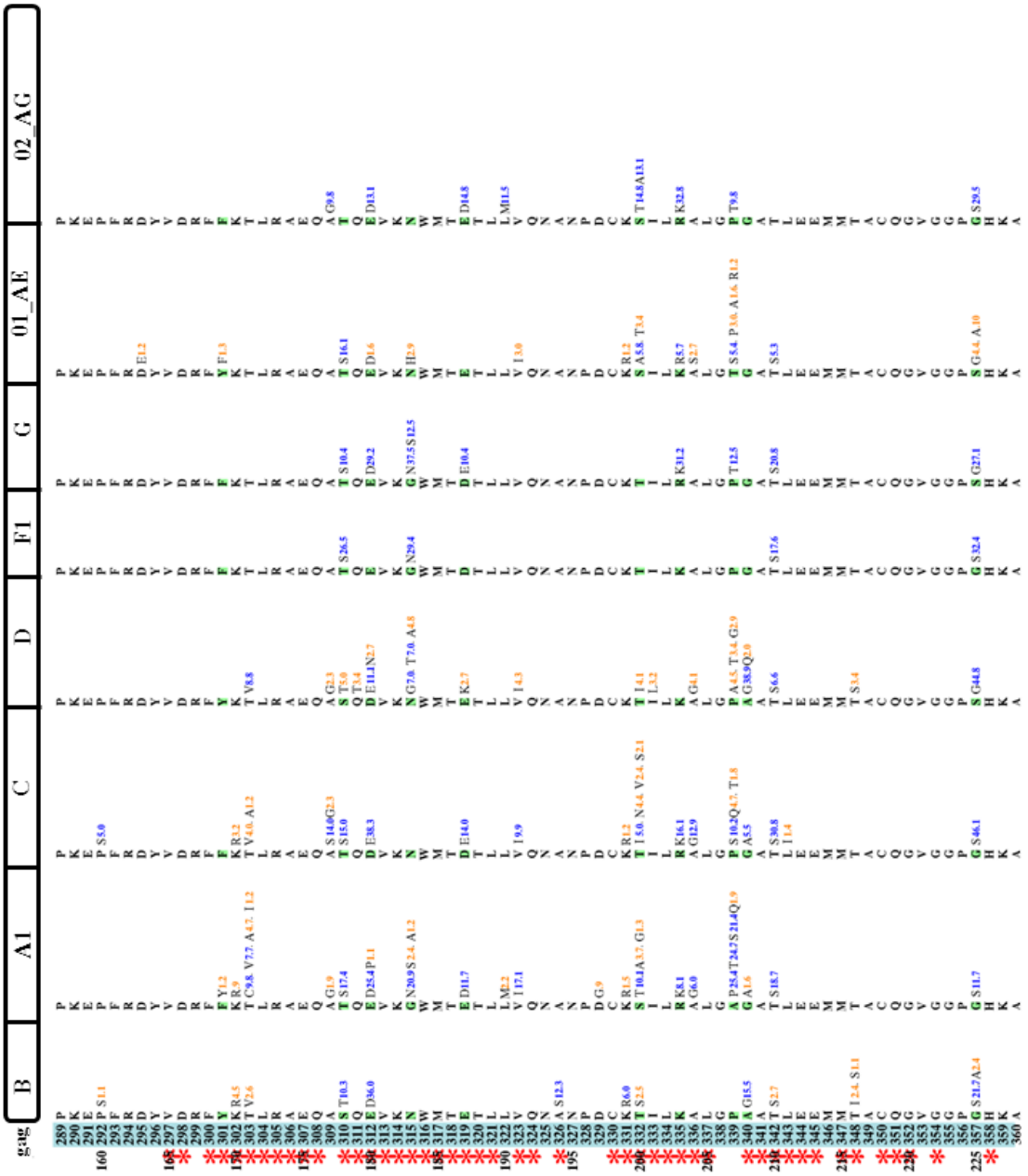


Figure S2.3 (continue)

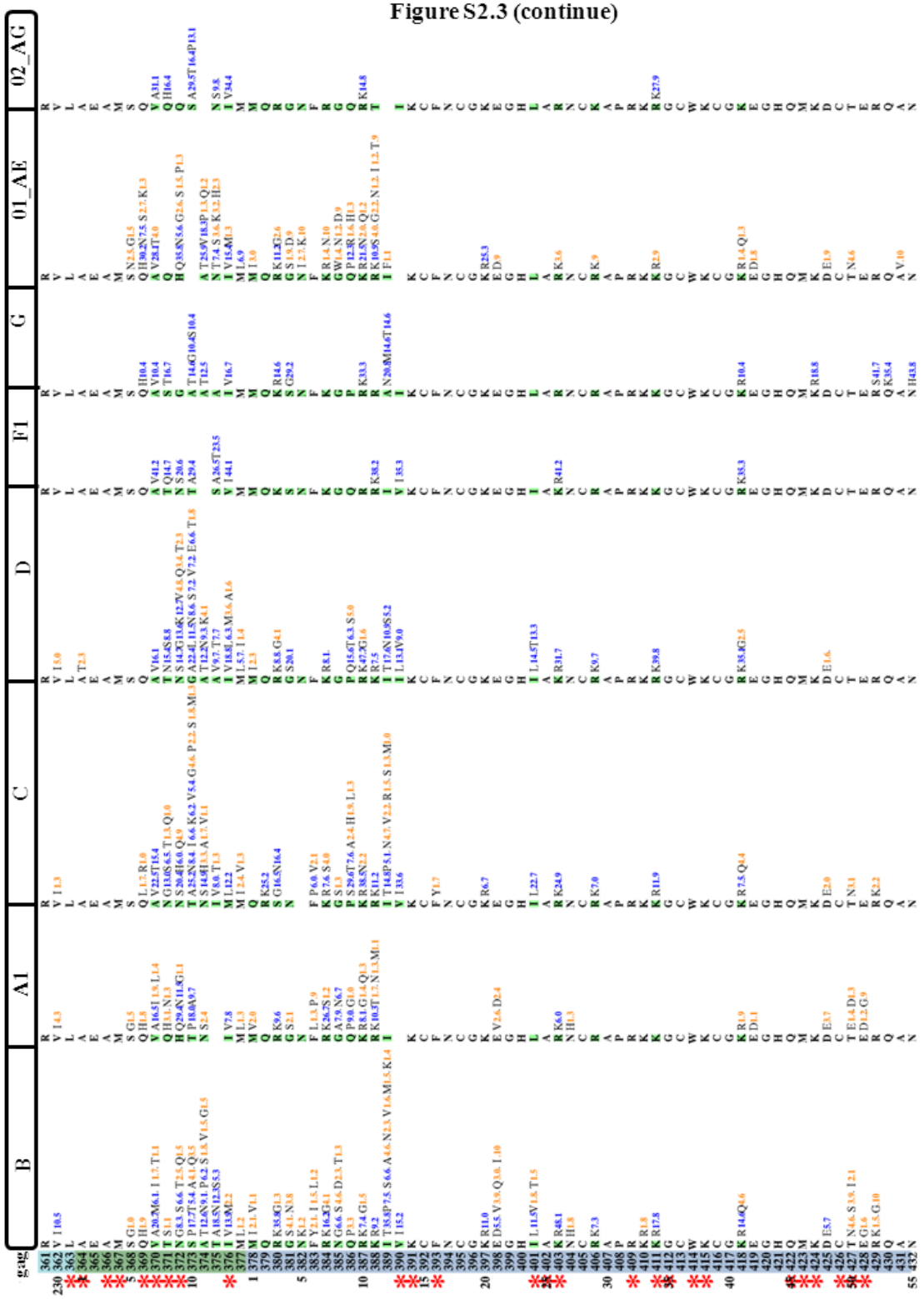
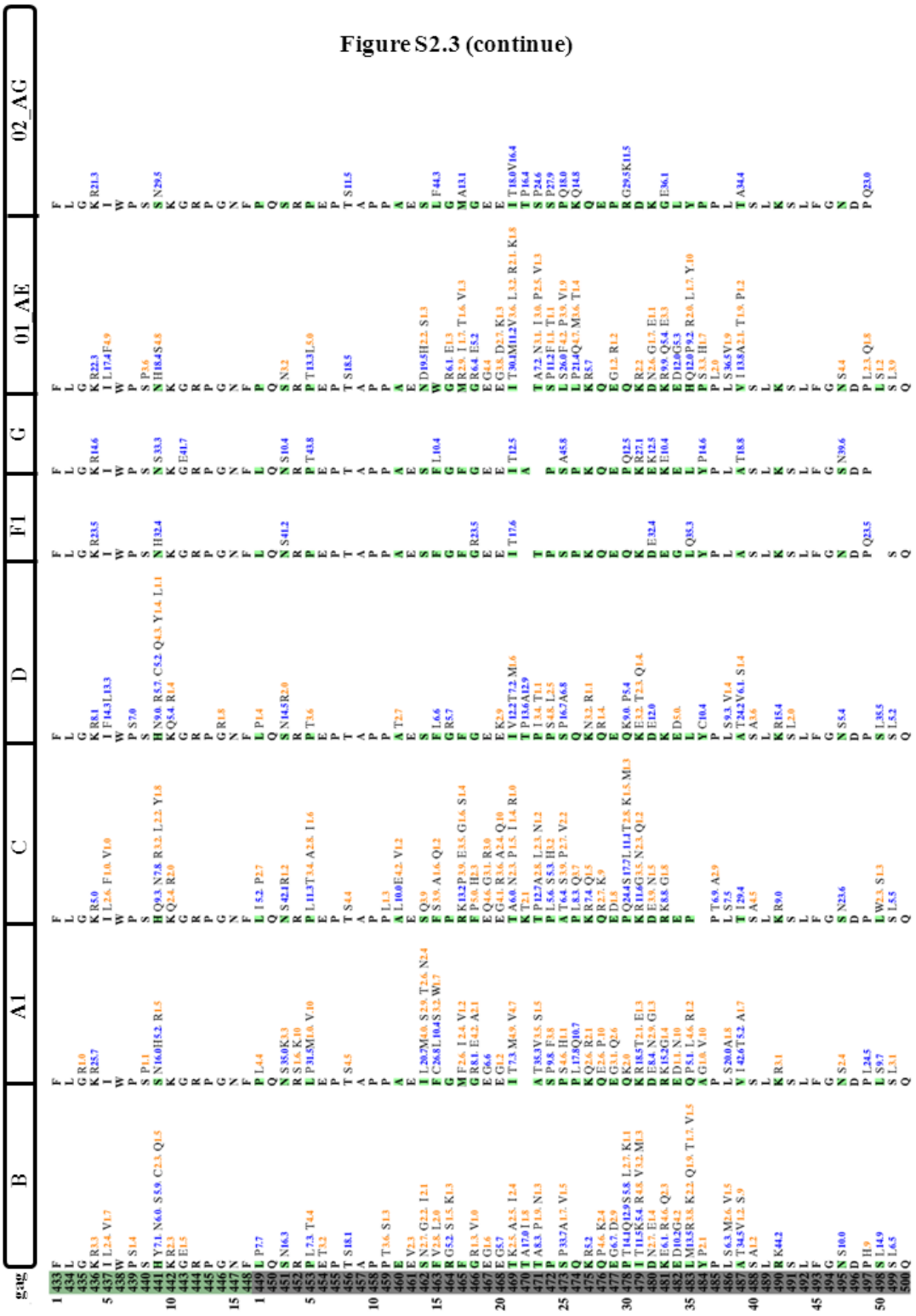


Figure S2.3 (continue)



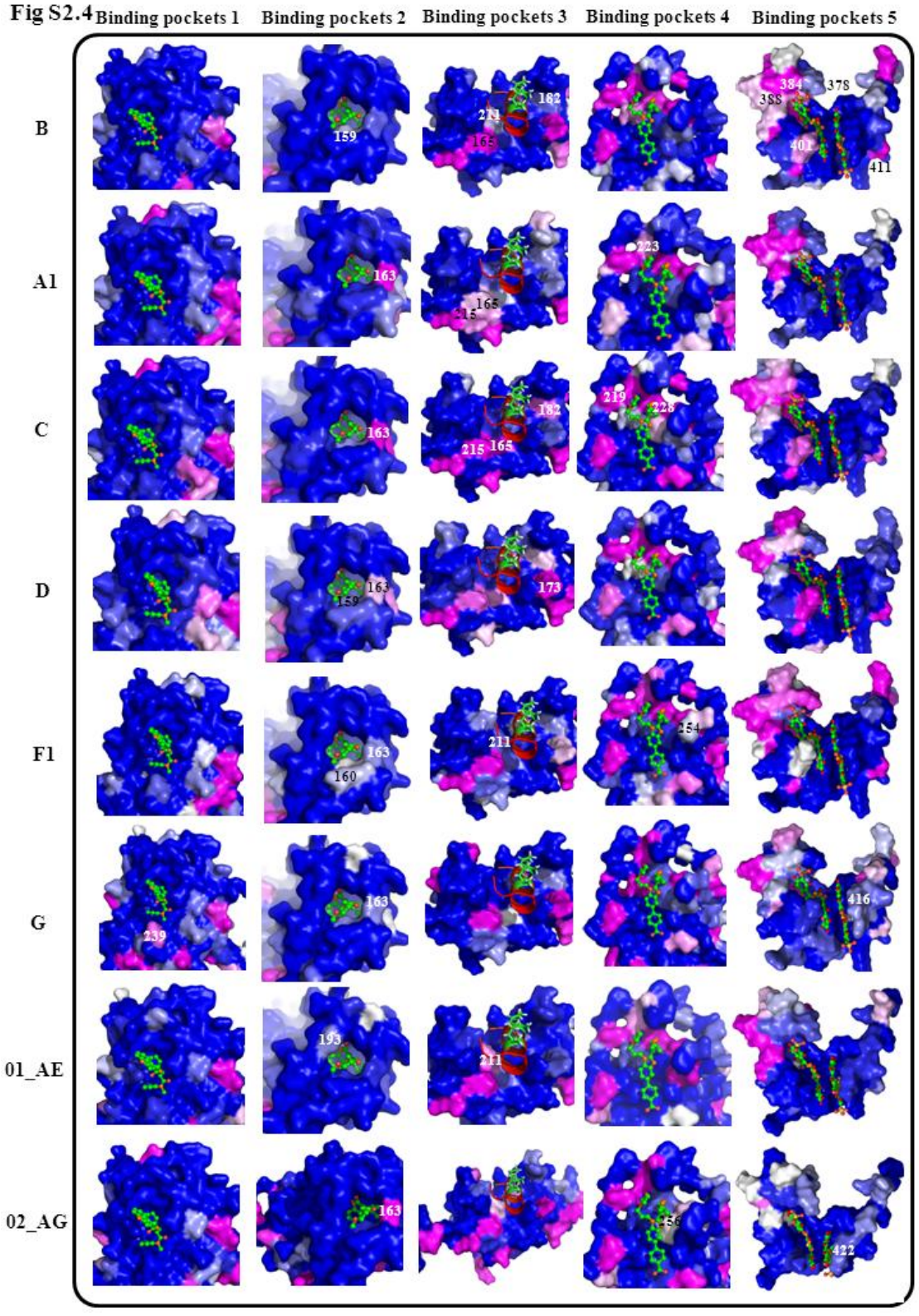
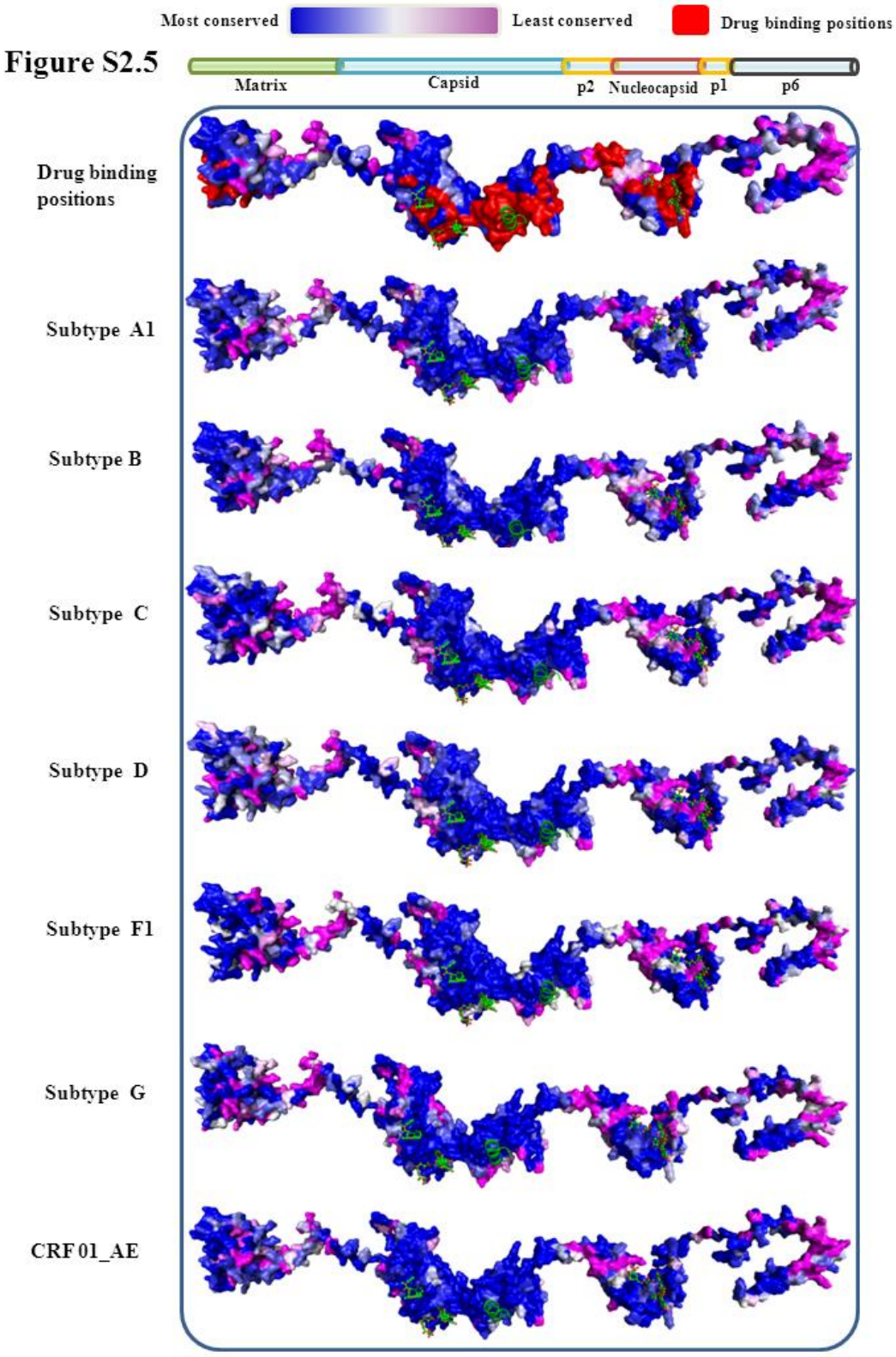


Figure S2.4: Surface representation of five drug binding pockets in 8 HIV-1 subtypes. For each subtype figure, surface spectrum colors indicate each position's CI, from the most conserved (blue CI = 0) to the least conserved positions (pink CI ≥ 0.1). Crystallized inhibitors are shown in sticks view inside their binding pockets. Visualization software: PyMOL V1.5 (http://www.pymol.org/).


Most conserved
Least conserved
Drug binding positions
Figure S2.5
Matrix
Capsid
p2
Nucleocapsid
p1
p6
Drug binding positions
Subtype A1
Subtype B
Subtype C
Subtype D
Subtype F1
Subtype G
CRF01_AE

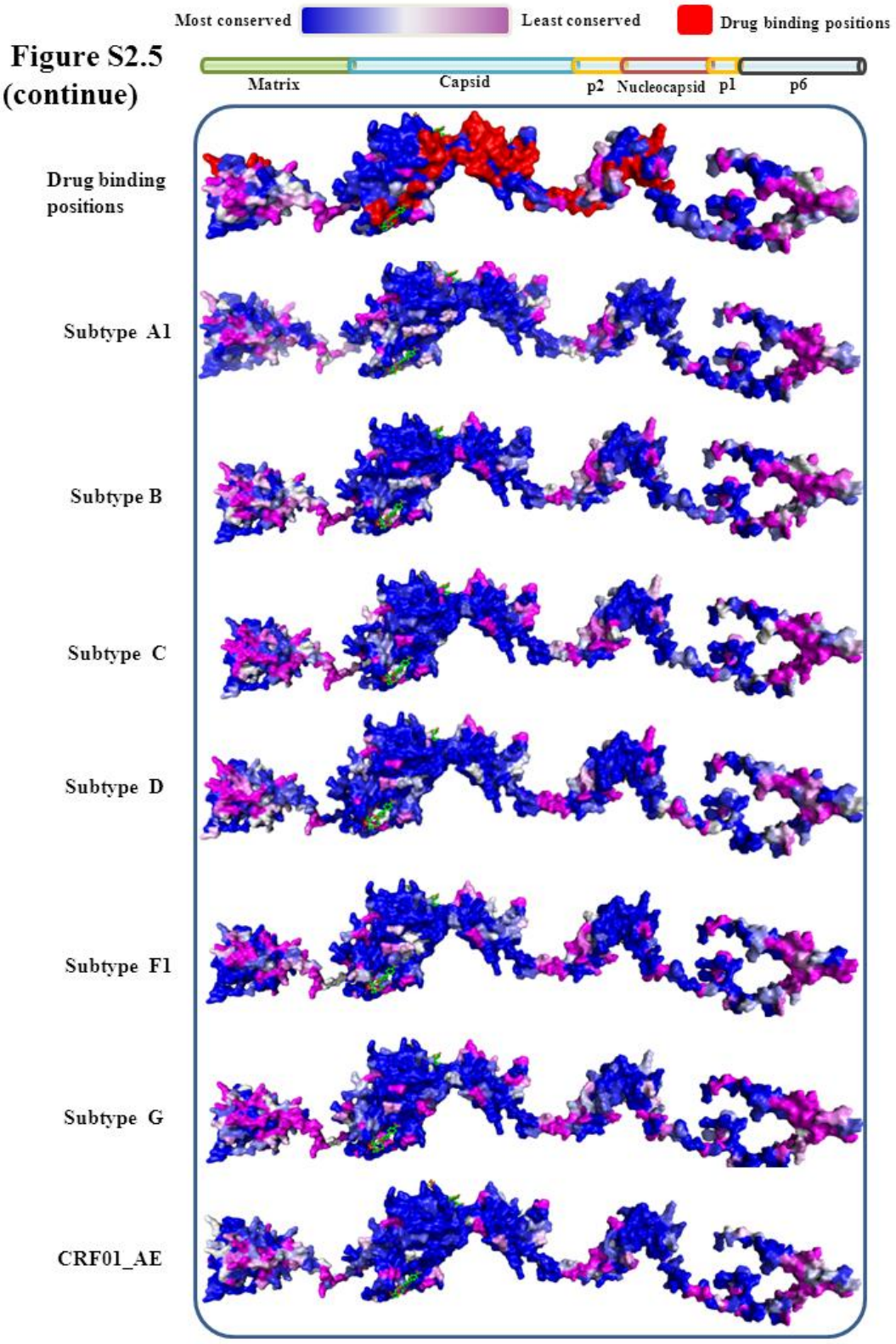


Figure S2.5: Surface representation of conservation index of full-length Gag in 8 HIV-1 subtypes. 4 Gag proteins (monomers) and 2 spacer peptides are annotated and displayed in schematic view at the top. Drug binding sites (red) are mapped onto HIV-1 Gag protein structures. For each subtype figure, surface spectrum colors indicate each position's CI, from the most conserved (blue CI = 0) to the least conserved positions (pink CI ≥ 0.1). Crystallized inhibitors are shown in sticks view

inside their binding pockets. Visualization software: PyMOL V1.5 (http://www.pymol.org/).

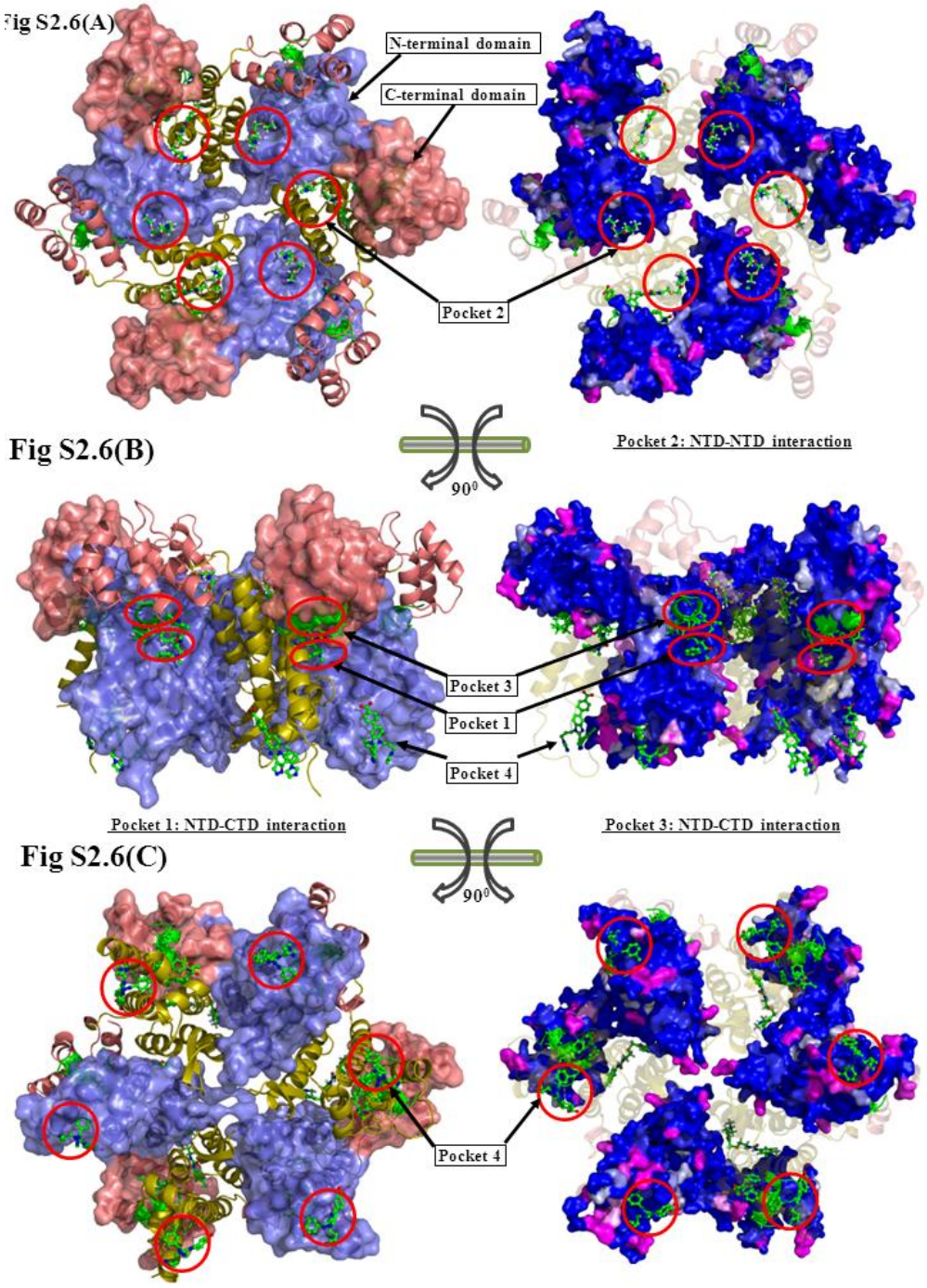


Figure S2.6: The structure of capsid hexamer superimposed with 8 crystallized inhibitors. The capsid hexamer structure is shown in the top, side and bottom views from Fig S1(A) to S1(C). The drug binding pockets are annotated in each figure. Three figures on the left side show the N-terminal domains (NTD, position: 1-146) and C-terminal domains (CTD, position: 151-231), colored blue and pink, respectively. Red circles indicate drug binding pockets, whose targets are within the interfaces of NTD-NTD, NTD-CTD or CTD-CTD interactions. Three figures on right side map the CIs to the structure and visualize the conservation of the drug binding pockets. It shows that drug binding pocket 1 in NTD is situated on the NTD-CTD

interaction interface, drug binding pocket 2 in NTD is situated on the NTD-NTD interaction interface, drug binding pocket 3 in CTD is situated on the NTD-CTD interaction interface, and drug binding pocket 4 is inside NTD. Note that drug binding pockets 1 and 3 are situated on the opposite sides of the same NTD-CTD interface. The superimposed crystallized inhibitors were mapped onto the capsid hexamer using PyMOL V1.5 (PDB: 3H4E). Visualization software: PyMOL V1.5 (http://www.pymol.org/). Inhibitor references are available in Additional file 1.

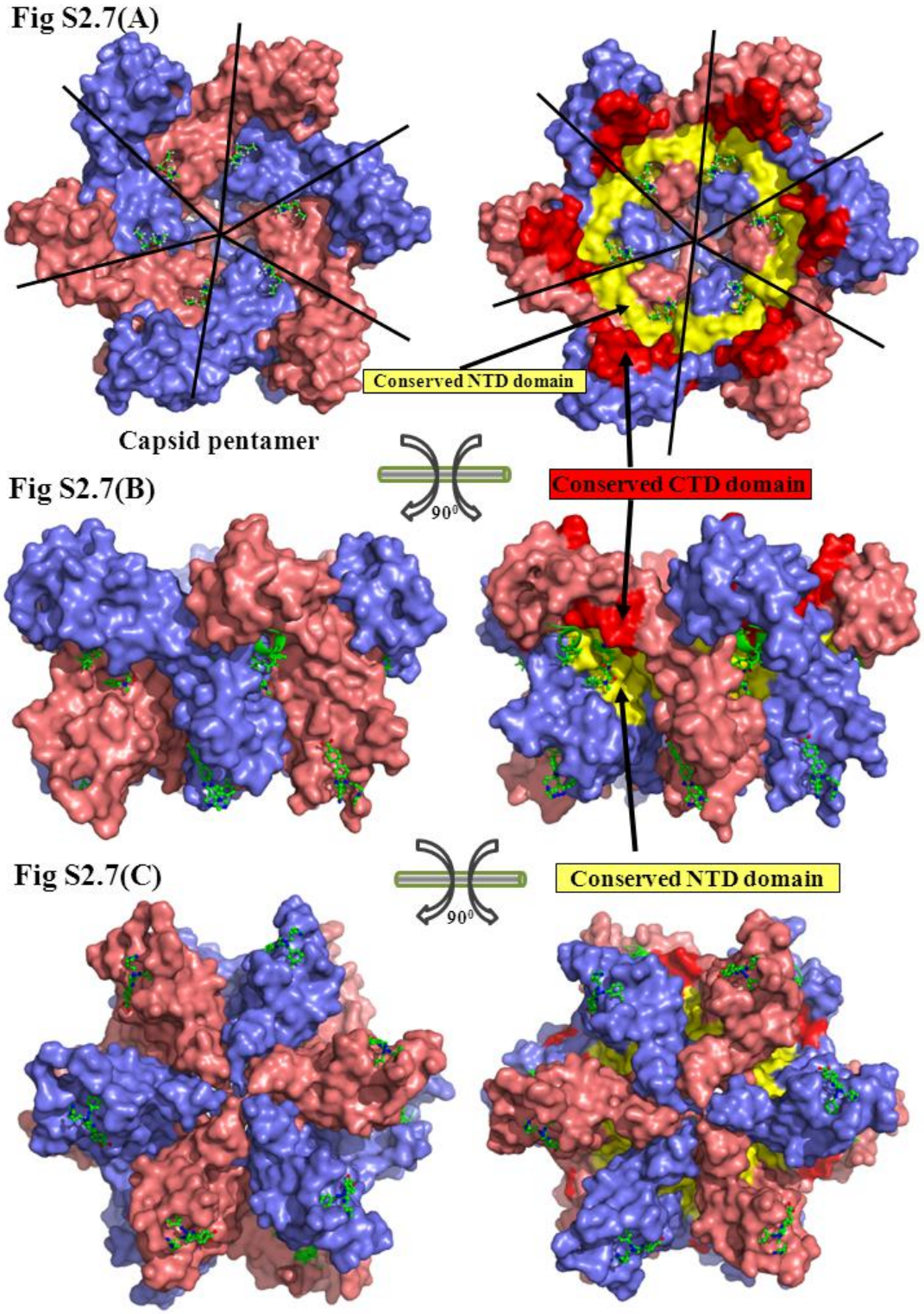


Figure S2.7: Conserved NTD and CTD domains in HIV-1 capsid. The capsid hexamer structure is shown in top (A), side (B) and bottom (C) views. Conserved NTD-NTD interaction domains are colored yellow (capsid positions: 30-70, Gag positions: 162-202). Conserved NTD-CTD interaction domains are colored red (capsid positions: 155-176, Gag positions: 287-308). PDB: 3H4E. Visualization software: PyMOL V1.5 (http://www.pymol.org/).

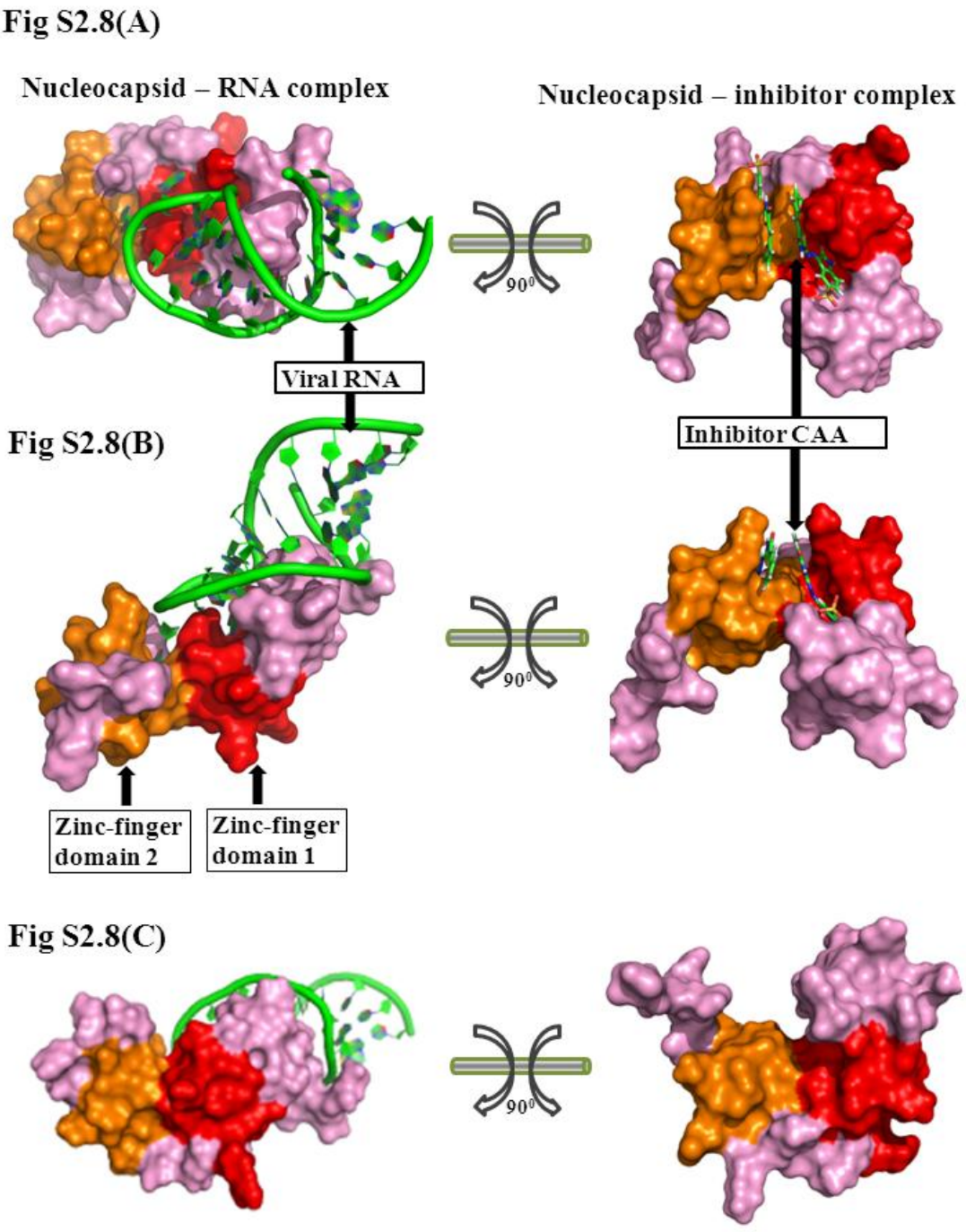


Figure S2.8: Conserved zinc-finger domains in HIV-1 nucleocapsid. The structures of nucleocapsid–RNA and nucleocapsid–inhibitor complexes are shown in top (A), side (B) and bottom (C) views. The first zinc-finger domain is colored red (nucleocapsid positions: 14-29, Gag positions: 389-404) and the second zinc-finger domain (nucleocapsid positions: 35-50, Gag positions: 410-425) are colored orange. PDB: 1A1T, 2M3Z. **Figure S2.7** and **Figure S2.8** visualize the conserved regions identified as three minimum conserved regions using our conservation analysis (**Figure 2.2**). Visualization software: PyMOL V1.5 (http://www.pymol.org/).

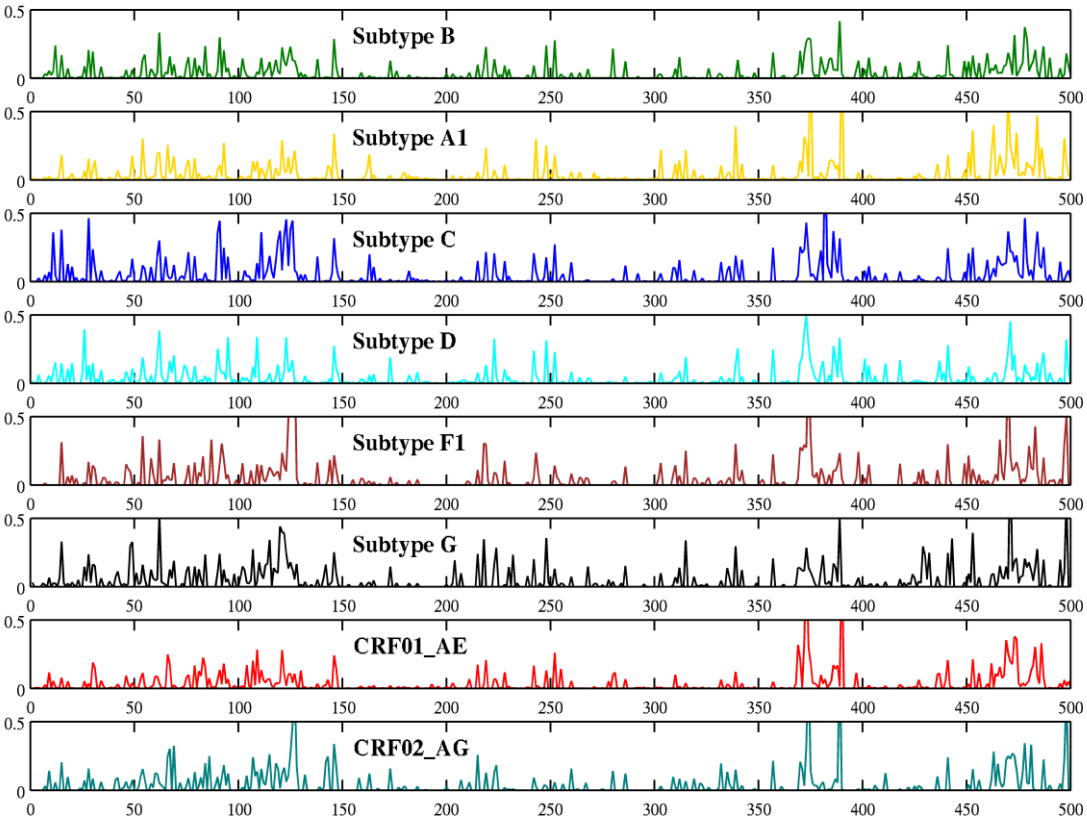


Figure S 2.9: Plots of CI values at 500 Gag positions across 8 HIV-1 subtypes and CRFs. Each subplot shows the results of A1, B, C, D, F1, G, CRF01_AE and CRF02_AG, respectively.

2.9 References

1. Engelman A, Cherepanov P. The structural biology of HIV-1: mechanistic and therapeutic insights. *Nat Rev Microbiol* 2012,**10**:279-290.
2. Waheed AA, Freed EO. HIV type 1 Gag as a target for antiviral therapy. *AIDS Res Hum Retroviruses* 2012,**28**:54-75.
3. Bocanegra R, Rodriguez-Huete A, Fuertes MA, Del Alamo M, Mateu MG. Molecular recognition in the human immunodeficiency virus capsid and antiviral design. *Virus Res* 2012,**169**:388-410.
4. Dau B, Holodniy M. Novel targets for antiretroviral therapy: clinical progress to date. *Drugs* 2009,**69**:31-50.
5. Salzwedel K, Martin DE, Sakalian M. Maturation inhibitors: a new therapeutic class targets the virus structure. *AIDS Rev* 2007,**9**:162-172.
6. Hemelaar J, Gouws E, Ghys PD, Osmanov S, Isolation W-UNfH, Characterisation. Global trends in molecular epidemiology of HIV-1 during 2000-2007. *AIDS* 2011,**25**:679-689.
7. Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, *et al.* Diversity considerations in HIV-1 vaccine selection. *Science* 2002,**296**:2354-2360.
8. Adamson CS, Sakalian M, Salzwedel K, Freed EO. Polymorphisms in Gag spacer peptide 1 confer varying levels of resistance to the HIV- 1 maturation inhibitor bevirimat. *Retrovirology* 2010,**7**:36.
9. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 2010,**27**:221-224.
10. Rose PP, Korber BT. Detecting hypermutations in viral sequences with an emphasis on G --> A hypermutation. *Bioinformatics* 2000,**16**:400-401.
11. Pineda-Pena AC, Faria NR, Imbrechts S, Libin P, Abecasis AB, Deforche K, *et al.* Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: Performance evaluation of the new REGA version 3 and seven other tools. *Infect Genet Evol* 2013.
12. Rhee SY, Liu TF, Kiuchi M, Zioni R, Gifford RJ, Holmes SP, *et al.* Natural variation of HIV-1 group M integrase: implications for a new class of antiretroviral inhibitors. *Retrovirology* 2008,**5**:74.
13. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, *et al.* The Protein Data Bank. *Nucleic Acids Res* 2000,**28**:235-242.
14. Zhang Z, Li Y, Lin B, Schroeder M, Huang B. Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics* 2011,**27**:2083-2088.
15. Brocchieri L, Karlin S. Conservation among HSP60 sequences in relation to structure, function, and evolution. *Protein Sci* 2000,**9**:476-486.
16. Blair WS, Pickford C, Irving SL, Brown DG, Anderson M, Bazin R, *et al.* HIV capsid is a tractable target for small molecule therapeutic intervention. *PLoS Pathog* 2010,**6**:e1001220.
17. Smith PF, Ogundele A, Forrest A, Wilton J, Salzwedel K, Doto J, *et al.* Phase I and II study of the safety, virologic effect, and pharmacokinetics/pharmacodynamics of single-dose 3-o-(3',3'-dimethylsuccinyl)betulinic acid (bevirimat) against human immunodeficiency virus infection. *Antimicrob Agents Chemother* 2007,**51**:3574-3581.
18. Nguyen AT, Feasley CL, Jackson KW, Nitz TJ, Salzwedel K, Air GM, *et al.* The prototype HIV-1 maturation inhibitor, bevirimat, binds to the CA-SP1 cleavage site in immature Gag particles. *Retrovirology* 2011,**8**:101.
19. Lu W, Salzwedel K, Wang D, Chakravarty S, Freed EO, Wild CT, *et al.* A single polymorphism in HIV-1 subtype C SP1 is sufficient to confer natural resistance to the maturation inhibitor bevirimat. *Antimicrob Agents Chemother* 2011,**55**:3324-3329.
20. Goudreau N, Lemke CT, Faucher AM, Grand-Maitre C, Goulet S, Lacoste JE, *et al.* Novel Inhibitor Binding Site Discovery on HIV-1 Capsid N-Terminal Domain by NMR and X-ray Crystallography. *ACS Chem Biol* 2013.
21. Bartonova V, Igonet S, Sticht J, Glass B, Habermann A, Vaney MC, *et al.* Residues in the HIV-1 capsid assembly inhibitor binding site are essential for maintaining the assembly-competent quaternary structure of the capsid protein. *J Biol Chem* 2008,**283**:32024-32033.
22. Lemke CT, Titolo S, von Schwedler U, Goudreau N, Mercier JF, Wardrop E, *et al.* Distinct effects of two HIV-1 capsid assembly inhibitor families that bind the same site within the N-terminal domain of the viral CA protein. *J Virol* 2012,**86**:6643-6655.

23. Schiffner T, Sattentau QJ, Dorrell L. Development of prophylactic vaccines against HIV-1. *Retrovirology* 2013,**10**:72.
24. Stephenson KE, Barouch DH. A global approach to HIV-1 vaccine development. *Immunol Rev* 2013,**254**:295-304.
25. Yufenyuy EL, Aiken C. The NTD-CTD intersubunit interface plays a critical role in assembly and stabilization of the HIV-1 capsid. *Retrovirology* 2013,**10**:29.
26. Zhang H, Curreli F, Zhang X, Bhattacharya S, Waheed AA, Cooper A, *et al.* Antiviral activity of alpha-helical stapled peptides designed from the HIV-1 capsid dimerization domain. *Retrovirology* 2011,**8**:28.
27. Goudreau N, Coulombe R, Faucher AM, Grand-Maitre C, Lacoste JE, Lemke CT, *et al.* Monitoring binding of HIV-1 capsid assembly inhibitors using (19)F ligand-and (15)N protein-based NMR and X-ray crystallography: early hit validation of a benzodiazepine series. *ChemMedChem* 2013,**8**:405-414.
28. Thomas JA, Gorelick RJ. Nucleocapsid protein function in early infection processes. *Virus Res* 2008,**134**:39-63.
29. Zentner I, Sierra LJ, Fraser AK, Maciunas L, Mankowski MK, Vinnik A, *et al.* Identification of a small-molecule inhibitor of HIV-1 assembly that targets the phosphatidylinositol (4,5)-bisphosphate binding site of the HIV-1 matrix protein. *ChemMedChem* 2013,**8**:426-432.
30. Alfadhli A, McNett H, Eccles J, Tsagli S, Noviello C, Sloan R, *et al.* Analysis of small molecule ligands targeting the HIV-1 matrix protein-RNA binding site. *J Biol Chem* 2013,**288**:666-676.
31. Frahm N, Kaufmann DE, Yusim K, Muldoon M, Kesmir C, Linde CH, *et al.* Increased sequence diversity coverage improves detection of HIV-specific T cell responses. *J Immunol* 2007,**179**:6638-6650.
32. Snoeck J, Fellay J, Bartha I, Douek DC, Telenti A. Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints. *Retrovirology* 2011,**8**:87.
33. Zentner I, Sierra LJ, Maciunas L, Vinnik A, Fedichev P, Mankowski MK, *et al.* Discovery of a small-molecule antiviral targeting the HIV-1 matrix protein. *Bioorg Med Chem Lett* 2013,**23**:1132-1135.
34. Fun A, Wensing AM, Verheyen J, Nijhuis M. Human Immunodeficiency Virus Gag and protease: partners in resistance. *Retrovirology* 2012,**9**:63.
35. Niedrig M, Gelderblom HR, Pauli G, Marz J, Bickhard H, Wolf H, *et al.* Inhibition of infectious human immunodeficiency virus type 1 particle formation by Gag protein-derived peptides. *J Gen Virol* 1994,**75 (Pt 6)**:1469-1474.
36. Ternois F, Sticht J, Duquerroy S, Krausslich HG, Rey FA. The HIV-1 capsid protein C-terminal domain in complex with a virus assembly inhibitor. *Nat Struct Mol Biol* 2005,**12**:678-682.
37. Bocanegra R, Nevot M, Domenech R, Lopez I, Abian O, Rodriguez-Huete A, *et al.* Rationally designed interfacial peptides are efficient in vitro inhibitors of HIV-1 capsid assembly with antiviral activity. *PLoS One* 2011,**6**:e23877.
38. Bhattacharya S, Zhang H, Debnath AK, Cowburn D. Solution structure of a hydrocarbon stapled peptide inhibitor in complex with monomeric C-terminal domain of HIV-1 capsid. *J Biol Chem* 2008,**283**:16274-16278.
39. Domenech R, Bocanegra R, Gonzalez-Muniz R, Gomez J, Mateu MG, Neira JL. Larger helical populations in peptides derived from the dimerization helix of the capsid protein of HIV-1 results in peptide binding toward regions other than the "hotspot" interface. *Biomacromolecules* 2011,**12**:3252-3264.
40. Dewan V, Liu T, Chen KM, Qian Z, Xiao Y, Kleiman L, *et al.* Cyclic peptide inhibitors of HIV-1 capsid-human lysyl-tRNA synthetase interaction. *ACS Chem Biol* 2012,**7**:761-769.
41. Vajdos FF, Yoo S, Houseweart M, Sundquist WI, Hill CP. Crystal structure of cyclophilin A complexed with a binding site peptide from the HIV-1 capsid protein. *Protein Sci* 1997,**6**:2297-2307.
42. Im YJ, Kuo L, Ren X, Burgos PV, Zhao XZ, Liu F, *et al.* Crystallographic and functional analysis of the ESCRT-I /HIV-1 Gag PTAP interaction. *Structure* 2010,**18**:1536-1547.
43. Li F, Goila-Gaur R, Salzwedel K, Kilgore NR, Reddick M, Matallana C, *et al.* PA-457: a potent HIV inhibitor that disrupts core condensation by targeting a late step in Gag processing. *Proc Natl Acad Sci U S A* 2003,**100**:13555-13560.
44. Vijay Baichwal HA, Brita Brown, Rena McKinnon, Kraig Yager,Vijay Kumar, David Gerrish, Mark Anderson and Robert Carlson. Anti-viral Characterization in vitro of a Novel

Maturation Inhibitor, MPC-9055. In: *Program & Abstracts of the 16th Conference on Retroviruses and Opportunistic Infections*. Montreal, Canada. ; 2009. Abstract 561.

45. Singh IP, Bodiwala HS. Recent advances in anti-HIV natural products. *Nat Prod Rep* 2010,**27**:1781-1800.
46. Kilgore N. RM, Zuiderhof M., Stanley D., Nitz T., Bullock P., Allaway G., Martin D. Characterization of PA1050040, a second generation HIV-1 maturation inhibitor. In: *IAS 2007, 4th IAS Conference On HIV Pathogenesis, Treatment and Prevention*. Sydney, Australia; 2007. Abstract MOPDX05.
47. Vijay Kumar DG, Christophe Hoarau, Kraig M. Yager, Harry Austin, Rena McKinnon, Brita Brown, Irene Dorweiler, Vijay Baichwal, Damon Papac, Chad Bradford, Scott Patton, Katrina Bulka, Lynn DeMie and Robert Carlson. Next Generation Orally Bioavailable HIV-1 Maturation Inhibitors. In: *239th ACS National Meeting & Exposition*. San Francisco,CA; 2010.
48. Blair WS, Cao J, Fok-Seang J, Griffin P, Isaacson J, Jackson RL, *et al.* New small-molecule inhibitor class targeting human immunodeficiency virus type 1 virion maturation. *Antimicrob Agents Chemother* 2009,**53**:5080-5087.
49. Waki K, Durell SR, Soheilian F, Nagashima K, Butler SL, Freed EO. Structural and functional insights into the HIV-1 maturation inhibitor binding pocket. *PLoS Pathog* 2012,**8**:e1002997.
50. Coric P, Turcaud S, Souquet F, Briant L, Gay B, Royer J, *et al.* Synthesis and biological evaluation of a new derivative of bevirimat that targets the Gag CA-SP1 cleavage site. *Eur J Med Chem* 2013,**62**:453-465.
51. Kortagere S, Madani N, Mankowski MK, Schon A, Zentner I, Swaminathan G, *et al.* Inhibiting early-stage events in HIV-1 replication by small-molecule targeting of the HIV-1 capsid. *J Virol* 2012,**86**:8472-8481.
52. Goudreau N, Lemke CT, Faucher AM, Grand-Maitre C, Goulet S, Lacoste JE, *et al.* Novel Inhibitor Binding Site Discovery on HIV-1 Capsid N-Terminal Domain by NMR and X-ray Crystallography. *ACS Chem Biol* 2013,**8**:1074-1082.
53. Urano E, Kuramochi N, Ichikawa R, Murayama SY, Miyauchi K, Tomoda H, *et al.* Novel postentry inhibitor of human immunodeficiency virus type 1 replication screened by yeast membrane-associated two-hybrid system. *Antimicrob Agents Chemother* 2011,**55**:4251-4260.
54. Fader LD, Landry S, Morin S, Kawai SH, Bousquet Y, Hucke O, *et al.* Optimization of a 1,5-dihydrobenzo[b][1,4]diazepine-2,4-dione series of HIV capsid assembly inhibitors 1: Addressing configurational instability through scaffold modification. *Bioorg Med Chem Lett* 2013,**23**:3396-3400.
55. Fader LD, Landry S, Goulet S, Morin S, Kawai SH, Bousquet Y, *et al.* Optimization of a 1,5-dihydrobenzo[b][1,4]diazepine-2,4-dione series of HIV capsid assembly inhibitors 2: Structure-activity relationships (SAR) of the C3-phenyl moiety. *Bioorg Med Chem Lett* 2013,**23**:3401-3405.
56. Kelly BN, Kyere S, Kinde I, Tang C, Howard BR, Robinson H, *et al.* Structure of the antiviral assembly inhibitor CAP-1 complex with the HIV-1 CA protein. *J Mol Biol* 2007,**373**:355-366.
57. Curreli F, Zhang H, Zhang X, Pyatkin I, Victor Z, Altieri A, *et al.* Virtual screening based identification of novel small-molecule inhibitors targeted to the HIV-1 capsid. *Bioorg Med Chem* 2011,**19**:77-90.
58. Tremblay M, Bonneau P, Bousquet Y, DeRoy P, Duan J, Duplessis M, *et al.* Inhibition of HIV-1 capsid assembly: optimization of the antiviral potency by site selective modifications at N1, C2 and C16 of a 5-(5-furan-2-yl-pyrazol-1-yl)-1H-benzimidazole scaffold. *Bioorg Med Chem Lett* 2012,**22**:7512-7517.
59. Goudreau N, Hucke O, Faucher AM, Grand-Maitre C, Lepage O, Bonneau PR, *et al.* Discovery and Structural Characterization of a New Inhibitor Series of HIV-1 Nucleocapsid Function: NMR Solution Structure Determination of a Ternary Complex Involving a 2:1 Inhibitor/NC Stoichiometry. *J Mol Biol* 2013.
60. Vercruysse T, Basta B, Dehaen W, Humbert N, Balzarini J, Debaene F, *et al.* A phenyl-thiadiazolylidene-amine derivative ejects zinc from retroviral nucleocapsid zinc fingers and inactivates HIV virions. *Retrovirology* 2012,**9**:95.
61. Mori M, Schult-Dietrich P, Szafarowicz B, Humbert N, Debaene F, Sanglier-Cianferani S, *et al.* Use of virtual screening for discovering antiretroviral compounds interacting with the HIV-1 nucleocapsid protein. *Virus Res* 2012,**169**:377-387.

62. Turpin JA, Song Y, Inman JK, Huang M, Wallqvist A, Maynard A, *et al.* Synthesis and biological properties of novel pyridinioalkanoyl thiolesters (PATE) as anti-HIV-1 agents that target the viral nucleocapsid protein zinc fingers. *J Med Chem* 1999,**42**:67-86.
63. Breuer S, Chang MW, Yuan J, Torbett BE. Identification of HIV-1 inhibitors targeting the nucleocapsid protein. *J Med Chem* 2012,**55**:4968-4977.
64. Avilov SV, Boudier C, Gottikh M, Darlix JL, Mely Y. Characterization of the inhibition mechanism of HIV-1 nucleocapsid protein chaperone activities by methylated oligoribonucleotides. *Antimicrob Agents Chemother* 2012,**56**:1010-1018.
65. Miller Jenkins LM, Ott DE, Hayashi R, Coren LV, Wang D, Xu Q, *et al.* Small-molecule inactivation of HIV-1 NCp7 by repetitive intracellular acyl transfer. *Nat Chem Biol* 2010,**6**:887-889.
66. Pannecouque C, Szafarowicz B, Volkova N, Bakulev V, Dehaen W, Mely Y, *et al.* Inhibition of HIV-1 replication by a bis-thiadiazolbenzene-1,2-diamine that chelates zinc ions from retroviral nucleocapsid zinc fingers. *Antimicrob Agents Chemother* 2010,**54**:1461-1468.
67. Shvadchak V, Sanglier S, Rocle S, Villa P, Haiech J, Hibert M, *et al.* Identification by high throughput screening of small compounds inhibiting the nucleic acid destabilization activity of the HIV-1 nucleocapsid protein. *Biochimie* 2009,**91**:916-923.
68. Srivastava P, Schito M, Fattah RJ, Hara T, Hartman T, Buckheit RW, Jr., *et al.* Optimization of unique, uncharged thioesters as inhibitors of HIV replication. *Bioorg Med Chem* 2004,**12**:6437-6450.
69. Schito ML, Goel A, Song Y, Inman JK, Fattah RJ, Rice WG, *et al.* In vivo antiviral activity of novel human immunodeficiency virus type 1 nucleocapsid p7 zinc finger inhibitors in a transgenic murine model. *AIDS Res Hum Retroviruses* 2003,**19**:91-101.
70. Sharmeen L, McQuade T, Heldsinger A, Gogliotti R, Domagala J, Gracheck S. Inhibition of the early phase of HIV replication by an isothiazolone, PD 161374. *Antiviral Res* 2001,**49**:101-114.
71. Rice WG, Baker DC, Schaeffer CA, Graham L, Bu M, Terpening S, *et al.* Inhibition of multiple phases of human immunodeficiency virus type 1 replication by a dithiane compound that attacks the conserved zinc fingers of retroviral nucleocapsid proteins. *Antimicrob Agents Chemother* 1997,**41**:419-426.
72. Witvrouw M, Balzarini J, Pannecouque C, Jhaumeer-Laulloo S, Este JA, Schols D, *et al.* SRR-SB3, a disulfide-containing macrolide that inhibits a late stage of the replicative cycle of human immunodeficiency virus. *Antimicrob Agents Chemother* 1997,**41**:262-268.
73. Huang M, Maynard A, Turpin JA, Graham L, Janini GM, Covell DG, *et al.* Anti-HIV agents that selectively target retroviral nucleocapsid protein zinc fingers without affecting cellular zinc finger proteins. *J Med Chem* 1998,**41**:1371-1381.
74. Mayasundari A, Rice WG, Diminnie JB, Baker DC. Synthesis, resolution, and determination of the absolute configuration of the enantiomers of cis-4,5-dihydroxy-1,2-dithiane 1,1-dioxide, an HIV-1NCp7 inhibitor. *Bioorg Med Chem* 2003,**11**:3215-3219.
75. Jenkins LM, Byrd JC, Hara T, Srivastava P, Mazur SJ, Stahl SJ, *et al.* Studies on the mechanism of inactivation of the HIV-1 nucleocapsid protein NCp7 with 2-mercaptobenzamide thioesters. *J Med Chem* 2005,**48**:2847-2858.
76. Goel A, Mazur SJ, Fattah RJ, Hartman TL, Turpin JA, Huang M, *et al.* Benzamide-based thiolcarbamates: a new class of HIV-1 NCp7 inhibitors. *Bioorg Med Chem Lett* 2002,**12**:767-770.
77. Rice WG, Turpin JA, Huang M, Clanton D, Buckheit RW, Jr., Covell DG, *et al.* Azodicarbonamide inhibits HIV-1 replication by targeting the nucleocapsid protein. *Nat Med* 1997,**3**:341-345.
78. Karlin S, Brocchieri L. Evolutionary conservation of RecA genes in relation to protein structure and function. *J Bacteriol* 1996,**178**:1881-1894.
79. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 1992,**89**:10915-10919.
80. Gong S, Blundell TL. Discarding functional residues from the substitution table improves predictions of active sites within three-dimensional structures. *PLoS Comput Biol* 2008,**4**:e1000179.
81. Valdar WS. Scoring residue conservation. *Proteins* 2002,**48**:227-241.
82. Capra JA, Singh M. Predicting functionally important residues from sequence conservation. *Bioinformatics* 2007,**23**:1875-1882.

Chapter 3

An integrated map of HIV genome-wide variation from a population perspective

"Minds are like flowers, they only open when the time is right"

— Stephen Richards

This chapter is adapted from my article:
Guangdi Li, Supinya Piampongsant, Nuno Rodrigues Faria, Arnout Voet, Andrea-Clemencia Pineda-Peña, Ricardo Khouri, Philippe Lemey, Anne-Mieke Vandamme, Kristof Theys. An integrated map of HIV genome-wide variation from a population perspective.

I proposed the idea, designed the software and drafted the manuscript. The improvement of the paper was supported with substantial help from Dr. Kristof Theys and Prof. Anne-Mieke Vandamme, as well as advices and corrections from other coauthors. I sincerely thank Fossie Ferreira and Jasper Edgar Neggers for technical assistance and valuable contributions to the analysis.

3.1 Summary

The HIV pandemic is characterized by extensive genetic variability, which has challenged the development of HIV drugs and vaccines. Although HIV genomes have been classified into different types, groups, subtypes and recombinants, a comprehensive study that maps HIV genome-wide diversity at the population level is still lacking to date. This study aims to characterize HIV genomic diversity in large-scale sequence populations, and to identify driving factors that shape HIV genome diversity. A total of 2996 full-length genomic sequences from 1705 patients infected with 16 major HIV groups, subtypes and circulating recombinant forms (CRFs) were analyzed along with structural, immunological and peptide inhibitor information. Average nucleotide diversity of HIV genomes was almost 50% between HIV-1 and HIV-2 types, 37.5% between HIV-1 groups, 14.7% between HIV-1 subtypes, 8.2% within individual HIV-1 subtypes and less than 1% within single patients. Along the HIV genome, diversity patterns and compositions of nucleotides and amino acids were highly similar across different groups, subtypes and CRFs. Current HIV-derived peptide inhibitors were predominantly derived from conserved, solvent accessible and intrinsically ordered structures in the HIV-1 subtype B genome. We identified these conserved regions in Capsid, Nucleocapsid, Protease, Integrase, Reverse transcriptase, Vpr and the GP41 N terminus as potential drug targets. In the analysis of factors that impact HIV-1 genomic diversity, we focused on protein multimerization, immunological constraints and HIV-human protein interactions. We found that amino acid diversity in monomeric proteins was higher than in multimeric proteins, and diversified positions were preferably located within human CD4 T cell and antibody epitopes. Moreover, intrinsic disorder regions in HIV-1 proteins coincided with high levels of amino acid diversity, facilitating a large number of interactions between HIV-1 and human proteins. This first large-scale analysis provided a detailed mapping of HIV genomic diversity and highlighted drug-target regions conserved across different groups, subtypes and CRFs. Our findings suggest that, in addition to the impact of protein multimerization and immune selective pressure on HIV-1 diversity, HIV-human protein interactions are facilitated by high variability within intrinsically disordered structures.

3.2 Introduction

As the causative agent of AIDS, the Human Immunodeficiency Virus (HIV) represents a worldwide threat to public health and the economy. The HIV pandemic is characterized by extensive genomic diversity caused by multiple factors including multiple zoonotic transmissions into human populations, high rates of viral evolution and recombination [1]. HIV has two major types, HIV-1 and HIV-2, which are further divided into groups, subtypes and recombinant forms. Globally, over 90% of HIV infections belong to HIV-1 group M viruses, which have been classified into 9 subtypes (A-D, F-H, J, K) and more than 50 circulating recombinant forms (CRFs) [1]. The high genetic diversity of the HIV genome has challenged the development of drugs and vaccines [2].

The HIV genome contains nine genes that encode fifteen viral proteins. Three major genes, *gag*, *pol* and *env*, code for structural proteins (Matrix, Capsid, Nucleocapsid, p6), enzymes (Protease, Reverse transcriptase (RT), Integrase) and envelope proteins (GP120, GP41), respectively. The remaining genes code for regulatory (Tat, Rev) and accessory proteins (Vif, Vpr, Vpu/Vpx, Nef) [3]. These viral proteins exhibit multiple functions and interact with various human proteins during the HIV life cycle [4, 5].

During the past three decades, many antiviral inhibitors have been designed to prevent HIV replication by targeting different viral proteins [6]. These anti-HIV peptides and small-molecule inhibitors either act by blocking active sites of viral enzymes or interrupting protein interactions [6]. For instance, the fusion inhibitor T20 (Enfuvirtide, Fuzeon), a peptide derived from the GP41 heptad repeat region, can efficiently inhibit viral entry by interrupting interactions between the GP41 helices [7]. For all existing drug classes, mutations in the HIV genome can cause drug resistance [8]. Therefore, inhibitors have been preferentially developed to target conserved regions of different viral proteins [9]. HIV genetic diversity also challenges the development of a global HIV vaccine [10]. While the vaccine trial STEP was unable to show preventive vaccination in subtype B infected cohorts [11], the Thai trial RV144 showed for the first time that prime-boost vaccination provided a modest efficacy in patients infected with CRF01_AE [12]. For vaccine and drug design, it

remains important to investigate the genomic diversity of different HIV groups, subtypes and CRFs at a population level.

Despite a large body of knowledge on different aspects of HIV pathogenesis, a large-scale analysis that reveals the genome-wide diversity within and between different HIV groups, subtypes and CRFs is still lacking. Although previous HIV genomic studies have reported subtype distribution, genetic variability, disease progression, evolutionary rate, positive selective pressure and the origin of HIV [11-27], most studies reported their findings using either reference genomes or small cohorts of less than 100 patients or sequences in a single subtype. HIV-1 subtype B which dominates infections in developed countries is the most studied subtype, largely due to historical reasons [28]. For instance, the adaptive evolution during acute infection was evaluated only in 11 individuals infected with HIV-1 subtype B [14]. In light of using HIV consensus sequences as vaccine candidates, an analysis on the genetic difference between consensus sequences and circulating strains was limited to subtypes B and C using less than 100 sequences [2], while other subtypes also prevail worldwide [29].

The last three decades have seen an accumulation of HIV data including full-length genomic sequences, protein crystal structures, HIV-human protein interactions, human T-cell epitope information and antiretroviral peptide inhibitors derived from the HIV genome. By integrating distinct but complementary sources of large-scale HIV datasets, this study aims to characterize HIV genome-wide diversity and to determine multiple factors that shape HIV genomic diversity.

3.3 Materials and Methods

HIV genome sequence dataset: In August 2013, we retrieved 3607 nucleotide genomic sequences of major HIV-1 and HIV-2 clades (HIV-2 group A and B, HIV-1 group N, O, P, subtype A1, B, C, D, F1, G, H, J, K, CRF01_AE, CRF02_AG) from the HIV Los Alamos database (www.hiv.lanl.gov/). The quality criteria for removing duplicates and sequences with hypermutations, stop codons, ambiguous nucleotides or subtype misclassification were described in [9]. The sequence dataset that fulfilled the quality criteria comprised 2996 genomic sequences, sampled from 1684 HIV-1 and 21 HIV-2 patients between 1982 and 2013. Information on genomic sequence datasets is summarized in **Table 3.1**.

Nucleotide genomic sequences were aligned using MUSCLE [30]. Protein regions encoded by their respective open reading frames (ORFs) were concatenated according to the reference strains (HIV-1: HXB2, HIV-2: BEN). For each HIV protein coding region, the translation of nucleotide to amino acid sequence alignments was optimized by our nucleotide to amino acid alignment toolbox. This toolbox maximizes amino acid matches, including in overlapping reading frames, based on the BLOSUM62 substitution matrix. Sequence alignments were further curated using Seaview v4.3 [31]. Our alignment toolbox and genomic sequences are available in Additional file 3.

Table 3.1: Information of HIV-1 and HIV-2 full-length genome sequence datasets

Type	HIV-1														HIV-2	
Group	M											N	O	P	A	B
Subtype/CRF	A1	B	C	D	F1	G	H	J	K	01_AE	02_AG					
Number of genome	159	1425	554	65	25	27	4	2	2	581	81	11	25	4	25	6
Number of patient	134	657	429	57	22	23	4	2	2	250	71	9	22	2	16	5
Average length in nucleotides*	8500	8600	8600	8500	8500	8600	8600	8600	8600	8500	8500	8500	8700	8600	8600	8600

*: Only the HIV coding regions are counted.

HIV-derived peptide inhibitor dataset: HIV-derived peptide inhibitors have their amino acid sequences derived from HIV proteins. We searched for English articles in PubMed published between January 1983 and September 2013 using the keywords "HIV peptide inhibitor", "HIV [protein name] peptide" and "HIV [protein name] inhibitor". References from primary studies, review articles and peptide design papers were also reviewed. If more than one peptide inhibitor were reported in one publication, only the most promising peptide inhibitors as indicated by the abstract of articles were collected. If data on the same inhibitors was reported by more than one publication, only the latest results were retained. **Table S 3.1** summarizes the 121 peptide inhibitors with corresponding information on peptide sequences, peptide-derived regions, target proteins, inhibitory activities and references.

PDB, HIV-human protein interaction, CD4/CD8/antibody epitope datasets: As of February 2014, we queried HIV PDB data from the RCSB Protein Data Bank using sequence search; PDB quality was then examined using PDBREPORT [32] (**Table 3.2**). We extracted HIV-human protein interactions (interaction type: physical interaction) from the NCBI HIV-1 human protein interaction database [33]. From the HIV Los Alamos database (http://www.hiv.lanl.gov/content/immunology/), we

extracted the human CD4 T cell and antibody epitopes in HIV-1. For human CTL/CD8 T cell epitopes, we included the best-defined CTL epitopes of the A-list described in [34] (**Table 3.3**).

Table 3.2: Summary of protein structures and PDB data for HIV-1 and HIV-2 proteins

Gene		***gag***						***pol***		
Protein		Matrix	Capsid	p2	Nucleocapsid	p1	p6	Protease	RT	Integrase
Number of units		3	5,6	1	1	1	1	2	2	4
HIV-1	AA length	132	231	14	55	16	52	99	560	288
	Multimer	1HIW #	3H4E	-	-	-	-	1A30	1N6Q	1K6Y
	Monomer	-	-	1U57	1A1T	-	2C55	-	-	-
HIV-2	AA length	135	229	17	52	14	64	99	559	292
	Multimer	-	2WLV	-	-	-	-	3S45	1MU2	-
	Monomer	2K4E	-	-	2E1X	-	-	-	-	3F9K

Gene		***vif***	***vpr***	***tat***	***rev***	***vpu/vpx****	***env***		***Nef***	
Protein		Vif	Vpr	Tat	Rev	Vpu/Vpx	GP120	GP41	Nef	
Number of units		2,3,4	1	2	2,6	1	3	3	2	
HIV-1	AA length	192	96	101	116	82	481	345	206	
	Multimer	-	-	-	3LPH	-	4NCO	2XRA	-	
	Monomer	4N9F	1M8L	1K5K	-	1VPU	-	-	4EMZ	
HIV-2	AA length	215	102	130	107	113	503	354	263	
	Multimer	-	-	-	-	-	-	-	-	
	Monomer	-	-	-	-	-	-	-	-	

#: PDB code from the RCSB Protein Data Bank; *: Vpu in HIV-1 and Vpx in HIV-2, -: either the data is not available or do not exist. For multerimeric HIV proteins, structures with different units can coexist such as pentamers and hexamers of Capsid [35], dimers and hexamers of Rev [36] and dimers, trimers and tetramers of Vif [37].

Table 3.3: **Summary of antibody, CD4+, CD8+ T cell epitope positions in the HIV-1 genome**

	Antibody epitope position	**CD4+ epitope position**	**CD8+ epitope position**
Matrix	20-31	1-107,118-132	11-44,74-101,124-132
Capsid	64-75	1-219	3-56,61-92,94-104,108-117,121-153,161-189,197-205, 217-231
p2		2-14	1-10
NC		1-55	28-36,50-55
p1		1-16	1-10
p6		1-43	33-41
Protease		53-70	3-11,30-42,57-66,68-90
RT	249-263,295-304,521-531	36-53,97-111,156-181,195-209,249-272,276-317,338-352,384-398,411-443	18-26,33-43,73-82,93-101,107-115,118-135, 137-166,173-187,202-210,244-252,260-279, 293-301,309-318,333-350,356-366,375-383, 392-401,436-457,495-505,520-528

Integrase		16-30,79-93,171-234,242-267	28-36,66-74,78-93,114-121,123-132,135-143,165-194,197-211,219-227,260-271
Vif		65-76,81-96	17-26,28-39,48-66,79-89,102-111,158-168
Vpr		32-96	29-42,48-67
Tat	47-60	17-55,64-80	30-49
Rev		9-56	14-23,57-75
Vpu		19-34	5-13,29-37
GP120	101-121,131-166,182-191,206-215,269-292,390-413,424-444,468-481	1-34,41-55,57-103,125-159,164-237,239-276,278-327,333-481	1-39,48-56,74-82,169-177,179-196,268-277,280-300,345-353,386-397
GP41	14-32,50-104,128-172	36-77,82-186,189-206,221-233,303-345	46-54,66-82,95-103,187-201,259-291,294-311,320-327,332-345
Nef	90-98	3-59,64-102,104-128,140-154,162-206	13-27,37-45,68-100,105-145,180-191

Protein secondary structure: For HIV-1 proteins (Rev, GP41) whose crystalized structures are not fully resolved in the PDB data, we used the sequence-based method PSIPRED V3.0 [38] to estimate protein secondary structures. For the other HIV-1 proteins with available PDB data (**Table 3.2**), we assessed protein secondary structures using both PSIPRED V3.0 [38] and 2Struc [39]. 2Struc is a software platform which integrates 8 PDB-based methods: DSSP_CONT, DSSP, KAKSI, PALSSE, P-SEA, STICKS, STRIDE and XTLSSTR [39]. Alpha-helix, beta-strand and random-coil structures were estimated using the majority voting of above 9 methods. Prediction similarities between these 9 methods are shown in **Figure S 3.11**.

Protein intrinsic disorder: Protein disordered regions are exploited by the virus to invade cellular host systems [40]; these regions are often structurally unstable without their partner molecules [41]. We estimated the intrinsically disordered structures of HIV-1 subtype B proteins using three software packages: MetaPrDOS [41], VSL2P [42] and PreDisorder v1.1 [43]. A disorder score (a numerical value between 0 and 1) of each amino acid position was estimated by 17 methods in these 3 software packages. An amino acid position was estimated as intrinsically disordered if its disorder score was above the cutoff value of 0.5 [41-43]. The intrinsically disordered positions were identified based on the majority voting of 17 methods. Prediction similarities between these 17 methods are shown in **Figure S 3.12**.

Table 3.4: Cutoffs for determining solvent exposed residues

Amino acid	A	R	N	D	C	Q	E	G	H	I
ASA(Å^2)	39.91 #	62.79	52.71	50.64	34.63	63.68	59.14	55.65	43.91	48.39
Amino acid	L	K	M	F	P	S	T	W	Y	V
ASA(Å^2)	55.76	52.23	63.85	50.23	45.29	37.80	46.01	76.61	71.12	47.58

#: For each amino acid, the cutoff is calculated using the 25% of the maximum ASA in all HIV-1 proteins, as described in [44].

Solvent accessible surface area: We estimated protein solvent accessible surface areas (ASA) using Chimera V1.6.1 [45] (default parameters). Provided with PDB data in **Table 3.2**, we calculated the ASA at each amino acid of all HIV-1 protein units. For each of the 20 amino acids, a distribution of its ASA scores over 15 HIV-1 proteins was obtained and the maximum ASA was identified therein. An amino acid at a specific position was considered buried if its ASA was lower than 25% of the maximum ASA for the corresponding amino acid [44] (**Table 3.4**).

Phylogenetic analysis: Our phylogenetic analysis was performed using 1384 nucleotide genomic sequences of 14 HIV groups and pure subtypes (thus excluding CRFs), obtained from the earliest sampling time (one sequence per patient). To prepare the alignment, we also removed ambiguous regions containing multiple insertions, deletions and hypervariable positions (HXB2 index: 1126-1182, 6866-7003, 7106-7154, 7773-7842, 7981-8032, 8897-9383). Maximum-likelihood phylogenetic trees were obtained using the multi-threaded FastTree V2.1 [46]. Our software parameters were set to 100 bootstrap replicates, the fully optimized GTR (generalized time-reversible) model, the continuous gamma distribution and the exhaustive nearest-neighbor interchange approach. The consensus phylogenetic tree with bootstrap supports was obtained using the seqboot tool in Phylip V3.69 (http://evolution.genetics.washington.edu/phylip.html).

Quantification of genetic diversity: Sequence diversity was calculated based on the pairwise nucleotide (NT) and amino acid (AA) comparisons [9, 47]. When calculating the amino acid diversity of HIV genome, we concatenated the amino acid sequences of 15 HIV protein coding regions in the full-length genome. Suppose the sequence dataset D contains L sequences with N positions, genetic diversity at position n is calculated by: $GD(D_n) = 1 - \frac{2}{L(L-1)} \sum_{i=1}^{L} \sum_{j=i+1}^{L} \delta(D_{n,i} = D_{n,j})$, where $D_{n,i}$ is the NT or AA form of the position n at the ith sequence in the dataset D, δ represents the Kronecker symbol, $\delta(D_{n,i} = D_{n,j})$ equals 1 if $D_{n,i}$ is identical to $D_{n,j}$; otherwise 0. Given the sequence dataset D, intra-clade genetic diversity $AGD(D)$ is defined as the average

genetic diversity of all positions: $AGD(D) = 1 - \frac{1}{N}\sum_{n=1}^{N}\frac{2}{L(L-1)}\sum_{i=1}^{L}\sum_{j=i+1}^{L}\delta(D_{n,i} = D_{n,j})$.

Suppose two sequence datasets D1 and D2 aligned with the same reference genome have the number of sequences L_1 and L_2 respectively. The inter-clade genetic diversity between D1 and D2 is defined as: $RGD(D1, D2) = 1 - \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L_1 \times L_2}\sum_{i=1}^{L_1}\sum_{j=1}^{L_2}\delta(D1_{n,i} = D2_{n,j})$. Furthermore, only positions for which less than 20% of sequences had gaps were considered and gaps were treated as missing data. Intra- and inter-clade genetic diversity was measured using one genomic sequence per patient, while intra-patient diversity was calculated using more than one genomic sequence sampled from individual patients. The Mann–Whitney U test was performed to compare the distributions of genetic diversity and a significant difference was identified if a p-value was less than 0.05. Our Matlab implementation of genomic diversity analysis is available in Additional file 3.

3.4 Results

Genome-wide diversity within and across HIV types, major groups and subtypes

We quantified the nucleotide and amino acid diversity of the HIV genome using 2996 full-length sequences sampled from 1705 patients **(Table 3.1**). The amino acid diversity was 53.8% (95% confidence interval (CI): 53.0-54.6%) between HIV-1 and HIV-2, 41.1% (CI: 25.6-54.3%) between HIV-1 groups, 18.0% (CI: 15.6-19.6%) between HIV-1 subtypes, 12.0% (CI: 8.6-14.4%) within HIV-1 subtypes and 1.1% (CI: 0.3-2.2%) within HIV-1 patients (**Figure 3.1**A). Similarly, nucleotide genomic diversity was found to be the highest when comparing HIV-1 and HIV-2 (mean: 48.32%, CI: 47.8-48.9%), followed by HIV-1 inter-group (37.5%, CI: 26.0-45.7%), HIV-1 inter-subtype (14.7%, CI: 12.2-15.8%), HIV-1 intra-subtype (8.2%, CI: 5.3-10.0%) and HIV-1 intra-patient diversity (0.6%, CI: 0.2-1.4%) (**Figure S 3.1**). As expected, the trend in HIV genomic diversity corresponds with the phylogenetic relationships between groups and pure subtypes in HIV-1 and HIV-2 (**Figure 3.1**B).

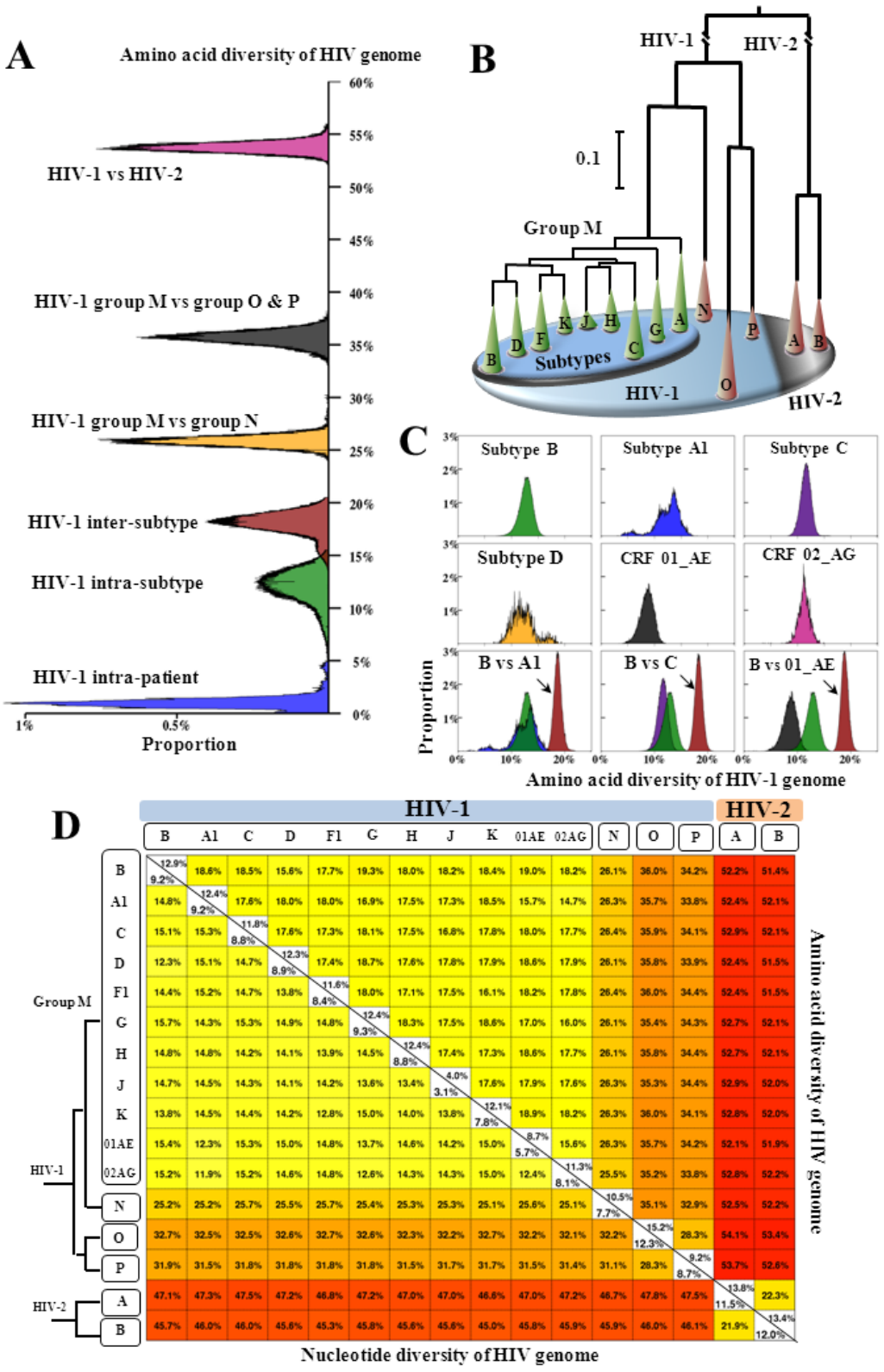


Figure 3.1: (A) Distribution plots of amino acid diversity in the HIV genome. The plots show the genomic diversity within HIV-1 infected patients (HIV-1 intra-patient, blue), within HIV-1 subtypes (HIV-1 intra-subtype, green), between HIV-1 subtypes (HIV-1 inter-subtype, red), between HIV-1 group M and group N (HIV-1 inter-group, yellow), between HIV-1 group M and group O/P (HIV-1 inter-group, black) and between HIV-1 and HIV-2 (pink). **Figure S 3.1** shows distribution plots of nucleotide genomic diversity.

(B) Maximum likelihood phylogenetic tree of HIV groups and pure subtypes. Green cones indicate HIV-1 subtypes in group M, while orange cones denote other HIV groups. All phylogenetic branches have bootstrap supports of more than 85% except one containing subtypes J, H and C. Branch lengths from the root to HIV-1 and HIV-2 are shortened for visualization purposes. SIV strains were not included in our phylogenetic tree. Software: FigTree V1.4.0 (http://tree.bio.ed.ac.uk/software/figtree/).
(C) Distribution plots of amino acid diversity in 6 major HIV-1 subtypes and CRFs (B, A1, C, D, CRF01_AE, CRF02_AG). X- and y-axes indicate the amino acid diversity and the proportions of sequence pairs, respectively. Six subplots in the first and second rows show the intra-subtype amino acid diversity of 6 HIV-1 subtypes and CRFs. Three subplots in the third row show the distribution of inter-subtype genomic diversity (B vs A1, B vs C, B vs 01_AE). One genomic sequence per patient (**Table 3.1**) was used for our analysis. **Figure S 3.2** shows the distribution of the other inter-clade genomic diversity.
(D) Average inter- and intra-clade genomic diversity of HIV-1 and HIV-2. The top right matrix demonstrates results for amino acid diversity, the bottom left matrix for nucleotide diversity. HIV subtypes and groups are shown on the left side of the matrix.

We next quantified genomic diversity within and between individual HIV clades. The distributions of inter-clade genomic diversity had mean values that were significantly higher than those of intra-clade genomic diversity (p-value < 0.05) (**Figure 3.1C, Figure S 3.2**). Within each HIV clade, amino acid diversity was consistently higher than nucleotide diversity (**Figure 3.1D**). CRF01_AE showed the lowest genomic diversity (nucleotide: 5.7%, amino acid: 8.7%) among the 10 HIV-1 subtypes with at least 10 sequences available (**Figure 3.1D**). Moreover, the estimated geographical distribution of HIV-1 genomic diversity (**Figure S 3.5**) showed a good agreement with the reported geographical distribution of HIV-1 subtypes [29].

Sequence variability was not uniformly distributed along the full-length HIV genome, but similar patterns were consistently observed in HIV group, subtype and CRF genomes at the nucleotide and amino acid levels (**Figure 3.2A, B**). Among all HIV proteins, Integrase was the most conserved protein (mean ± deviation: 4.5±1.1%), while GP120 varied the most (21.3±2.5%) (**Table 3.5**). Pairwise comparisons of genetic diversity between subtype B and the other clades identified conserved regions in the Capsid, Nucleocapsid, Protease, RT, Integrase, Vpr and the N terminus of GP41 (**Figure 3.2C**). Despite the different degrees of sequence diversity along the full-length genome, the nucleotide and amino acid compositions were comparable across the 16 group and subtype genomes (**Figure 3.3A, B**).

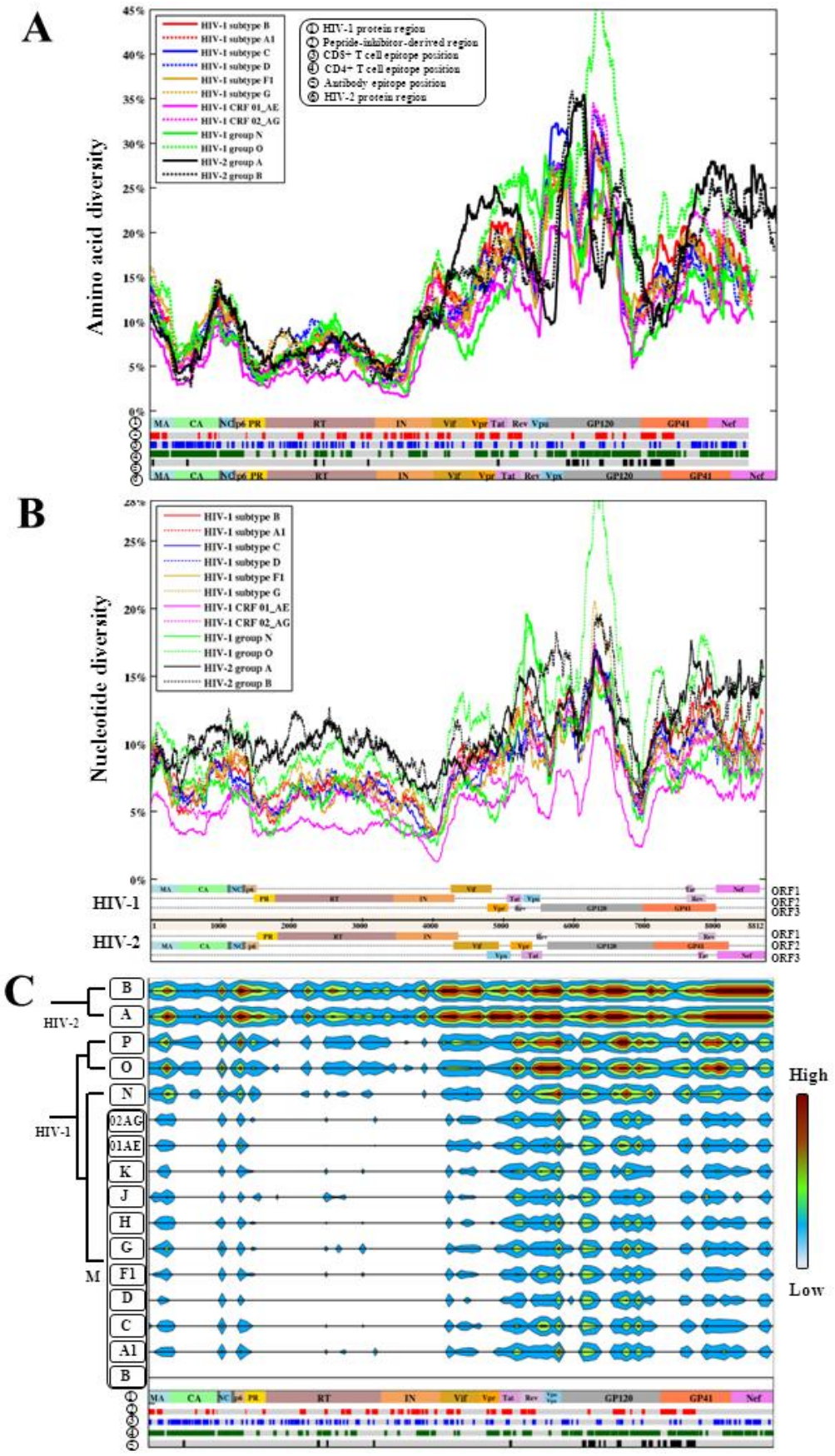


Figure 3.2: (A) Amino acid diversity along the full-length HIV genome using the sliding windows (window size: 100AA; also see the plots of exact diversity values in

Figure S 3.3). Each colored plot shows the density of amino acid diversity for one HIV group, subtype or CRF genome, indicated by the figure legend. Six layers are shown beneath the plots: (1) HIV-1 protein regions (HXB2 reference) are concatenated and shown with abbreviated names (e.g. MA: matrix); (2) peptide-inhibitor-derived region; (3) CD8+ T cell epitope position; (4) CD4+ T cell epitope position; (5) antibody epitope position; (6) HIV-2 protein region (BEN reference).
(B) Nucleotide diversity along the full-length HIV genome using sliding windows (window size: 300 nucleotides; also see the plots of exact diversity values in **Figure S 3.4**). Each colored plot shows the density of nucleotide diversity for one HIV group, subtype or CRF genome, indicated by the figure legend. Annotated HIV-1 and HIV-2 reference genomes are shown beneath; each track contains one open reading frame (ORF). Long terminal regions in the HIV genome are not shown.
(C) Contour map of inter-clade amino acid diversity between HIV-1 subtype B and the other HIV genomes. Inter-clade amino acid diversity was calculated by a sliding window of 30 amino acids over the HIV genome (low: ≤1 AA difference, high: ≥25 AA differences). Five colored layers beneath the contour map are annotated in (A).

Table 3.5: Average AA diversity of viral proteins within individual HIV clades (%)

	Clade #	MA	CA	NC	p6	PR	RT	IN	Vif	Vpr	Tat	Rev	Vpu/Vpx*	GP120	GP41	Nef
HIV-1	**Subtype A1**	13.07	7.3	7	19.89	5.59	6.44	4.76	13.63	9.86	18.14	16.83	23.04	23.34	13.38	14.64
	Subtype B	12.9	4.95	10.83	14.99	8.21	6.04	4.91	14.71	11.29	19.32	18.34	20.04	23.93	15.59	17.79
	Subtype C	16.11	5.77	9.39	15.52	6.15	5.84	4.45	10.61	10.64	15.37	16.10	20.21	22.89	14.19	14.96
	Subtype D	14.78	5.08	10.85	13.16	7.68	6.98	4.80	11.57	11.74	17.89	15.87	19.40	23.58	13.61	14.42
	Subtype F1	13.69	5.71	10.07	16.72	8.84	6.42	4.76	11.04	9.26	17.01	15.70	18.27	20.24	13.39	14.02
	Subtype G	17.33	5.54	7.40	17.21	6.60	7.30	4.17	14.36	10.33	15.51	18.39	19.25	23.87	12.98	14.15
	Subtype H	18.81	3.61	8.48	17.65	4.55	5.95	5.71	15.46	9.11	19.28	19.80	18.31	21.51	12.33	16.34
	CRF01_AE	10.06	3.00	4.92	13.37	4.34	3.75	2.39	9.32	7.97	13.57	13.34	12.80	17.40	9.62	11.78
	CRF02_AG	13.57	5.69	3.78	13.63	5.56	5.67	4.06	13.44	7.70	13.58	15.49	16.42	23.37	12.09	15.45
	Group N	10.52	6.40	7.06	10.01	5.93	6.78	3.09	11.02	4.87	12.15	12.77	35.01	21.43	9.15	12.64
	Group O	14.52	7.45	9.89	22.17	8.20	6.51	5.77	16.17	12.16	21.61	21.25	27.12	28.65	19.35	17.96
HIV-2	**Group A**	10.15	4.67	11.82	11.02	7.14	6.89	6.78	12.74	22.55	22.59	22.75	10.40	18.55	16.47	21.98
	Group B	12.14	3.54	13.15	15.4	5.76	6.4	6.19	12.71	12.66	20.44	15.82	11.45	21.20	15.29	20.59

#: Only HIV groups or subtypes with more than 2 genome sequences are listed (**Table 3.1**). *: Vpu in HIV-1 and Vpx in HIV-2.

Multiple factors shape HIV-1 genomic diversity

We next evaluated three potential factors (protein multimerization, immunological constraints, HIV-human protein interactions) that shaped the HIV genomic diversity. Firstly, we calculated the average diversity at amino acid positions of the 15 HIV-1 proteins (**Figure 3.3C**). For every HIV-1 group, subtype and CRF, the average amino

acid diversity was significantly higher in the monomeric proteins (Nucleocapsid, Vpr, Vpu, p6) than in the multimeric proteins (Matrix, Capsid, Protease, RT, Integrase, Vif, Tat, Rev, GP120, GP41, Nef) (p-value < 0.01) (**Table 3.6**). This suggests that the protein multimerization imposes a constraint on the HIV-1 sequence variability.

Secondly, we evaluated the amino acid variation in the known CD4 T cell, CD8 T cell and antibody epitopes (**Table 3.3**). By measuring the diversity of 3066 amino acid positions, we identified 919 (30%) variable positions with amino acid diversity above 12.9% (the average amino acid diversity within subtype B) using 657 subtype B genomic sequences. Univariate analysis showed that these variable positions were preferably located within antibody epitopes (OR 1.43, CI: 1.15-1.79, Fisher's exact test, p-value = 0.0015) and CD4 T cell epitopes (OR 1.73, CI: 1.18-2.96, p-value = 0.0438), but not within CD8 T cell epitopes (OR 1.11, CI: 0.82-1.51, p-value = 0.498) (**Figure 3.2**A).

Table 3.6: Comparison of average genetic diversity of HIV monomeric and multimeric proteins

HIV-1 clade	A1	B	C	D	F1	G	H
Monomers	14.6%	14.4%	14.4%	14.3%	14.1%	14.1%	13.9%
Multimers	12.7%	13.1%	12.5%	12.5%	12.3%	12.3%	12.3%
P-value [#]	5.4E-7	1.3E-5	3.8E-10	1.9E-11	5.2E-11	6.6E-11	4.9E-10
HIV-1 clade	J	K	01_AE	02_AG	N	O	P
Monomer	12.9%	13.0%	12.7%	12.5%	12.7%	13.0%	12.8%
Multimer	11.3%	11.3%	11.1%	11.1%	11.1%	11.4%	11.1%
P-value	1.1E-9	2.0E-9	6.3E-12	8.2E-11	9.1E-13	4.0E-14	8.9E-14

#: Mann-Whitney U-test. Monomeric proteins: Nucleocapsid, Vpr, Vpu, p6. Multimeric proteins: Matrix, Capsid, Protease, RT, Integrase, Vif, Tat, Rev, GP120, GP41, Nef.

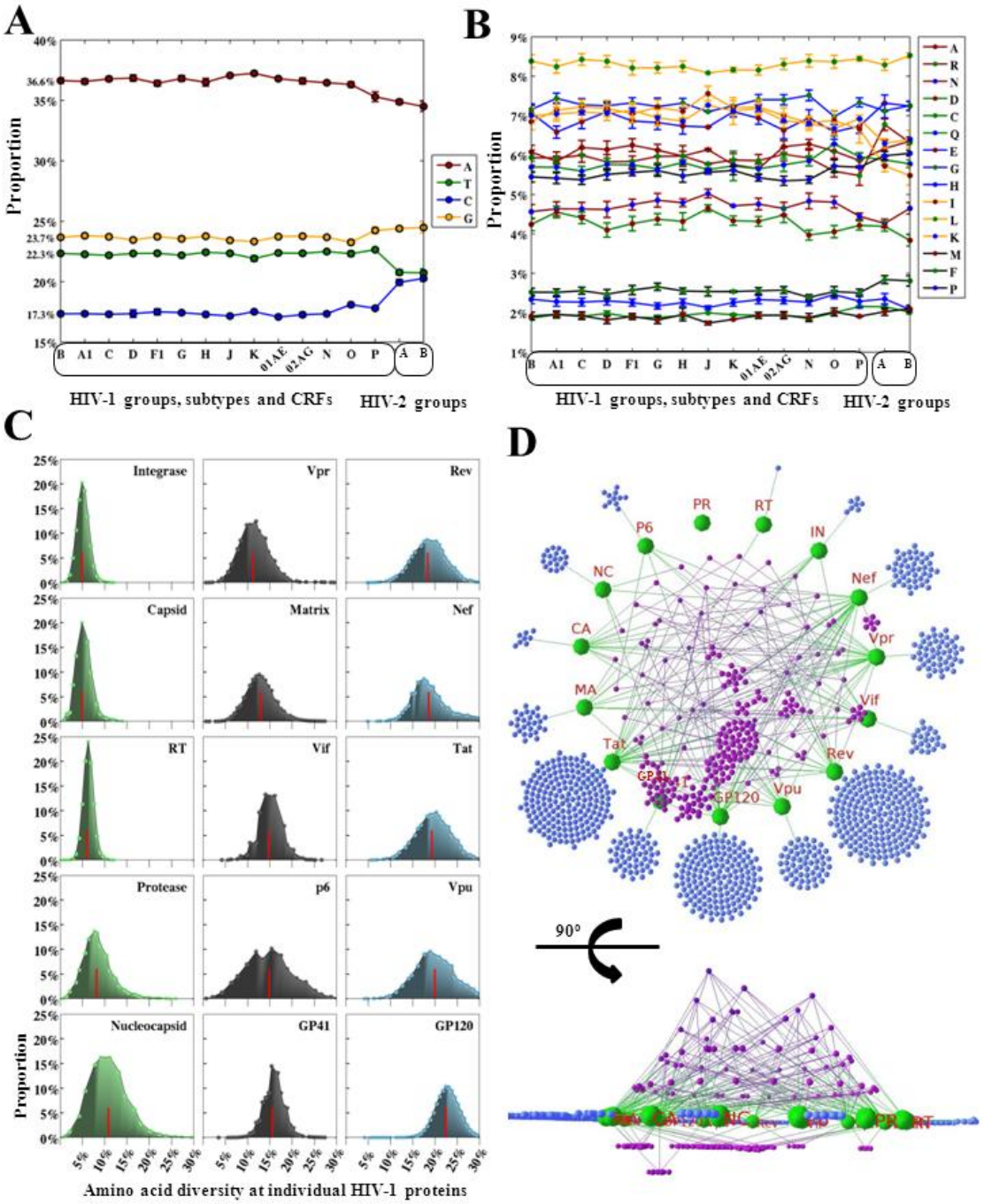


Figure 3.3: (A) Nucleotide composition for HIV-1 and HIV-2. X-axis represents the HIV groups, subtypes and CRFs. Y-axis shows the average proportions of nucleotides (A, T, C, G) using the HIV genomic sequence datasets (one sequence per patient, **Table 3.1**).

(B) Amino acid composition for HIV-1 and HIV-2. X-axis represents HIV groups, subtypes and CRFs. Y-axis shows the average proportions of amino acids using the HIV protein sequence datasets (one sequence per patient, **Table 3.1**).

(C) Distribution plots of amino acid genetic diversity for 15 HIV-1 subtype B proteins. Each subplot demonstrates a viral protein. X- and y-axes indicate the amino acid diversity and the proportions of amino acid diversity, respectively. Red lines inside distribution plots indicate mean values of amino acid diversity at individual proteins.

(D) Top and side views of 3D HIV-human protein interaction networks. HIV-1 proteins with protein names annotated are indicated by green spheres. Human proteins

that interact with only one HIV-1 protein are indicated by blue spheres in the outer circle (one human protein one sphere). Human proteins that interact with more than one HIV-1 protein are indicated by purple spheres above the plane of HIV-1 proteins. The height of the layers above the plane indicates the number of HIV proteins that a human protein interacts with. Below, human proteins are clustered if they interact with a set of more than one HIV-1 protein. Abbreviation names have been described in the abbreviation list. Visualization software: Geomi V2.0 (http://sydney.edu.au/engineering/it/~visual/geomi2/).

Thirdly, we mapped 1352 interactions between 1052 human and 15 HIV-1 proteins using the HIV-human protein interaction dataset (**Figure 3.3D**, see Materials). The following three observations support the hypothesis that the amino acid diversity of HIV-1 proteins is associated with HIV-human protein interactions. (1) Univariate analysis showed that HIV-1 proteins with higher amino acid diversity interact with more human proteins (Pearson's coefficient = 0.74, p-value = 0.0017). Polynomial regression analysis further identified a second-order model that fitted the correlation between these two variables (**Figure 3.4**A, adjusted R-squared: 0.82). (2) Intrinsically disordered structures in HIV-1 proteins can interact with multiple interaction partners [40]. Univariate analysis showed a significant correlation between the average amino acid diversity and the average disorder scores of HIV-1 proteins (Pearson's coefficient=0.64, p-value = 0.015, **Figure 3.4B**). (3) The levels of HIV-human protein interactions clustered according to the functional roles of the HIV-1 proteins, which have different functional roles and requirements for interactions with human proteins (**Figure 3.4C**). HIV regulatory proteins (Tat, Rev) and envelope proteins (GP120, GP41) had the largest number of interactions with different human proteins (568 for the regulatory proteins, 322 for the envelope proteins), while viral enzymes had the least number of interactions (**Figure 3.4C**). The average amino acid diversity of envelope proteins (20.4%) and regulatory proteins (18.8%) was higher than that of accessory proteins (16.0%), structural proteins (9.0%) and viral enzymes (5.9%) (**Figure S 3.6**). Our findings suggest that HIV-1 proteins with higher genetic diversities have larger intrinsically disordered structures and interact with more human proteins.

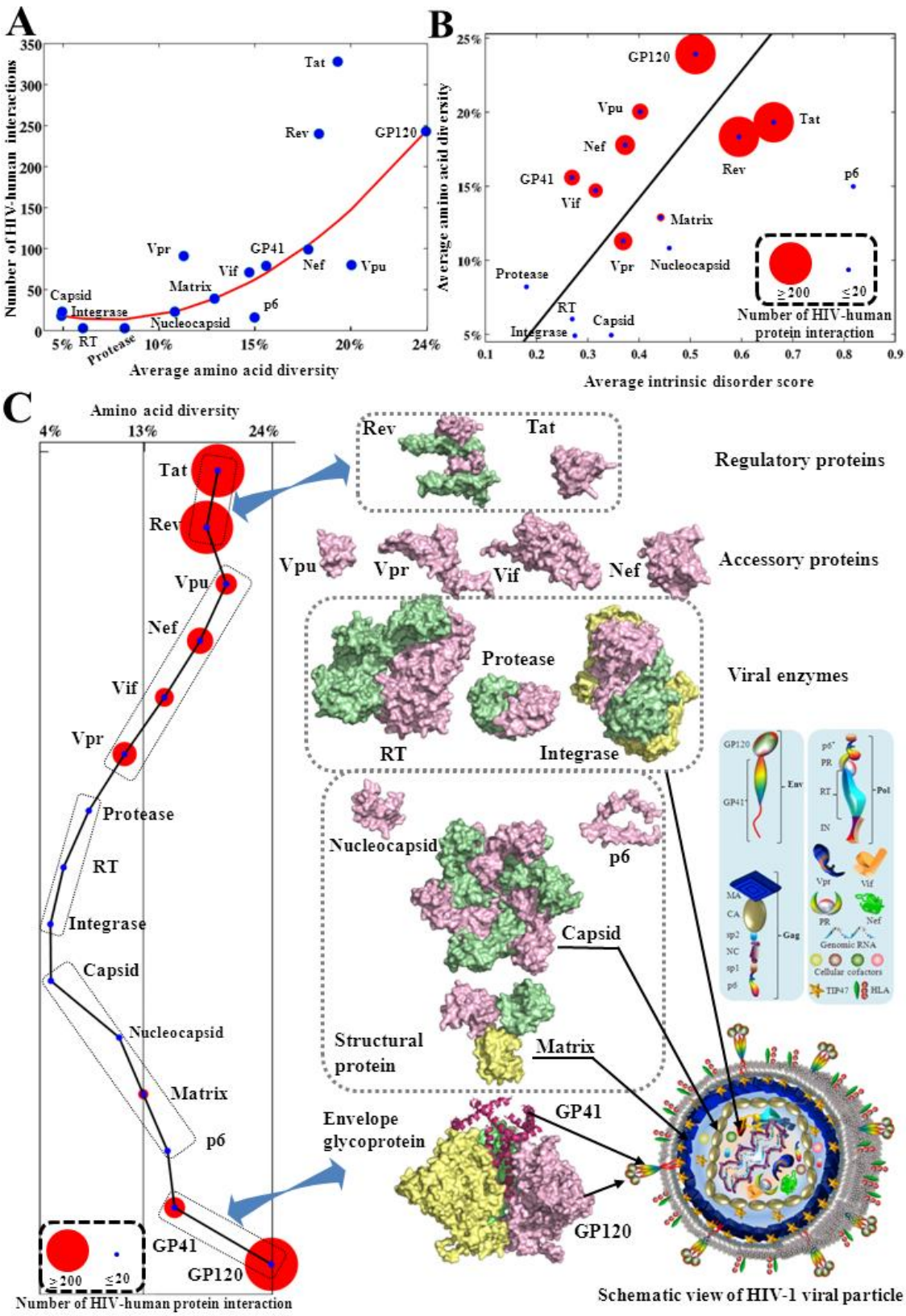


Figure 3.4: (A) Plot of polynomial regression between the HIV-1 protein diversity (x-axis) and the number of HIV-human protein interactions (y-axis). The second-order model is: $y = 8346x^2 - 1223x + 57.96$, (adjusted R-squared: 0.82, root-mean-square error: 42.31).

(B) Plot of average protein disorder score and average amino acid diversity in HIV-1 proteins. Red circles indicate the number of HIV-human protein interactions at individual viral proteins, for visualization purpose, scaled between 20 and 200 interactions (proteins with fewer than 20 interactions are scaled to the same size as those with 20, proteins with more than 200 interactions are scaled to the same size as those with 200). Average amino acid diversities of HIV-1 proteins are calculated using subtype B sequences (one genomic sequence per patient, **Table 3.1**).
(C) Clustering of HIV-1 proteins and schematic view of HIV-1 viral particle. On the left, each colored circle represents a viral protein positioned according to the clusters of protein functions. The size of each red circle indicates the number of HIV-human protein interactions involving each HIV-1 protein (see (B)). On the right, the schematic view of mature viral particle is visualized at the bottom with annotations indicated in the inserted figure legend. Above, surface representations show the structures of HIV-1 proteins that are grouped according to their functional roles. Different units in HIV-1 multimeric proteins are indicated with different colors and HIV-1 monomeric proteins are colored pink. HIV-1 protein structures are scaled according to their precise protein sizes for direct comparison. Visualization: PyMOL V1.5 (http://www.pymol.org/).

Peptide inhibitors are mainly derived from conserved subtype B genomic regions

We investigated the 121 HIV-derived peptide inhibitors reported between 1993 and 2013 (**Table S 3.1**). **Figure 3.5**A illustrates the GP41 structure and the GP41-derived region of T20 as an example of HIV-derived peptide inhibitors. Peptide inhibitors had on average a length of 25 AAs (range: 3 to 73), a charge of +0.27 at pH 7.2 and a molecular weight of 2953 g/mol. Most common amino acids in these peptide inhibitors were leucine, glutamic acid and isoleucine (**Figure S 3.7**). Comparisons between the 121 peptide sequences and the consensus sequences of 16 HIV group, subtype and CRF genomes showed the highest sequence similarity with subtype B (79.8%) (**Figure 3.5**B). Aspartic acid to asparagine (25.7%) was the most common amino acid substitution between the consensus subtype B sequence and the peptide inhibitor sequences (**Figure 3.5**C).

We characterized peptide-derived regions in the subtype B genome. Of the 894 amino acid positions from which the 121 peptide inhibitors were derived, 41.2% were located in helix structures and 60.2% displayed less than 5% genetic diversity in the subtype B genome. Forty-two inhibitors had IC_{50} or EC_{50} values less than 1 μM and were derived from 249 amino acid positions in the HIV-1 genome (**Table S 3.1**). In the subtype B genome, these 249 positions displayed significantly lower amino acid

diversity compared to the genome-wide diversity (**Figure 3.5**D, 10.1% vs 12.9%, p-value = 0.019), and were likely to be from conserved (amino acid diversity < 5%, OR: 1.43 (1.09-1.88), p-value = 0.016), solvent exposed (OR: 2.47 (1.88-3.24), p-value = 3.9E-11) and intrinsically ordered structures (disorder score < 0.4, OR: 1.75 (1.21-2.51), p-value = 0.0019) (**Figure 3.5**E).

Integrated findings from our analyses on HIV-1 genomic diversity, HIV-derived peptide inhibitors and protein structures are visualized in **Figure 3.6**. The HIV genomic sequence datasets and our toolbox developed for data visualization, genomic diversity analysis and HIV genomic alignment are freely available in Additional file 3.

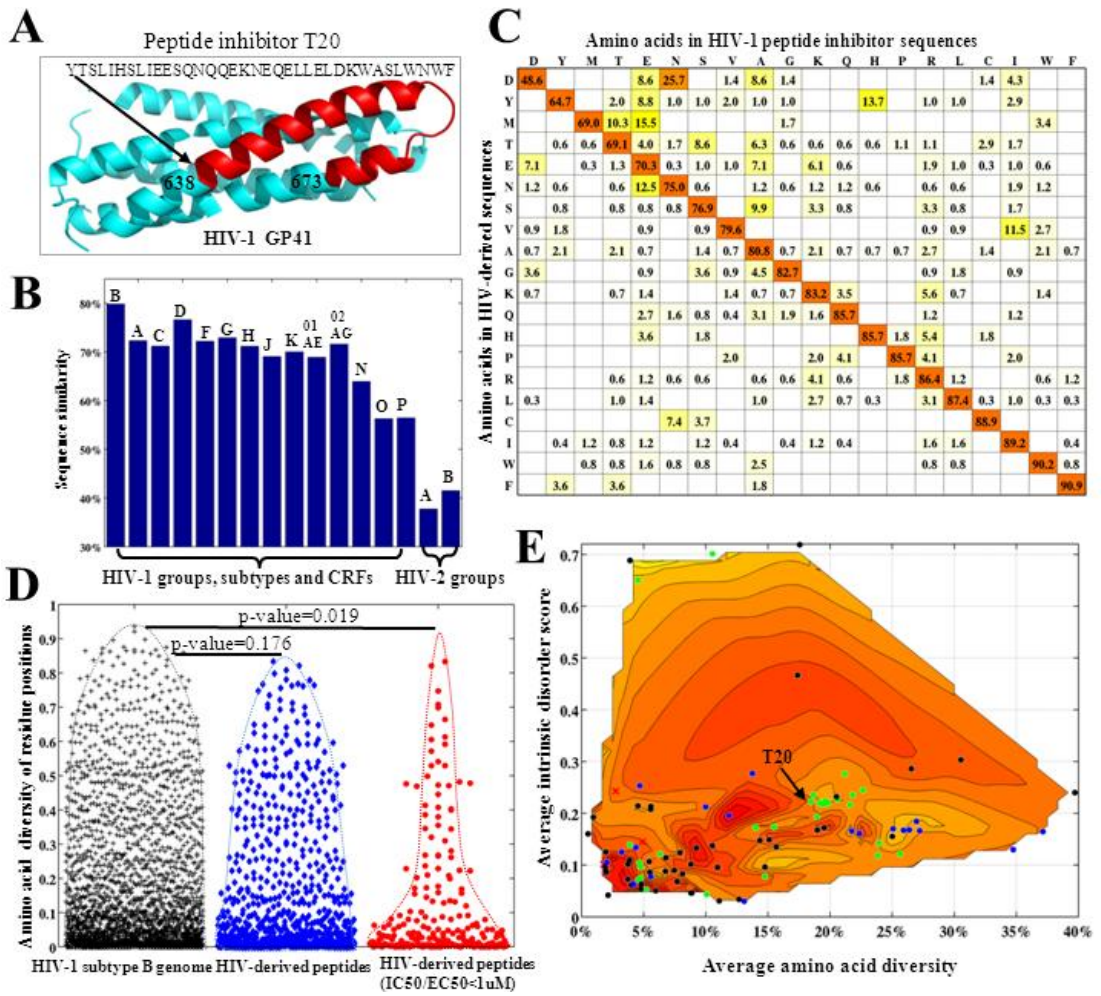


Figure 3.5: (A) Cartoon representation of GP41 structure. The red structure indicates the region from which peptide inhibitor T20 was derived (PDB: 3H01).
(B) Bar plot of sequence similarities between peptide inhibitor sequences and the sequences of HIV-derived regions in the consensus genome of different HIV clades. X-axis presents the HIV groups, subtypes and CRFs. Y-axis shows the sequence similarity between peptide inhibitor sequences and the sequences of HIV-derived regions in the consensus genomes of HIV groups, subtypes or CRFs.

(C) Amino acid replacements between peptide inhibitor sequences and HIV-derived regions in the subtype B genome. The percentage values (%) are colored using heat maps.
(D) Distribution (bee-swarm) plots of amino acid diversity in the full-length subtype B genome (black crosses), peptide-derived regions (blue diamonds) and peptide-derived regions of those inhibitors whose IC50/EC50 are less than 1 μM (red circles). Each shape represents the amino acid diversity at one protein position. Two-sample Kolmogorov-Smirnov tests were performed to compare diversity distributions (significance level: 0.05).
(E) Plot of amino acid diversity (x-axis), disorder score (y-axis) and solvent accessible surface area of peptide-inhibitor-derived regions (contour map, darker red indicates larger accessible surface areas). GP41 inhibitor T20 is also annotated. For individual peptide inhibitors, the average amino acid diversity, disorder score and solvent accessible surface areas are shown in **Figure S 3.8, Figure S 3.9 and Figure S 3.10**, respectively.

3.5 Discussion and conclusions

To our knowledge, this study provides the first large-scale analysis that investigates the genomic variability of 16 major groups, subtypes and CRFs in HIV-1 and HIV-2. While previous studies have reported the diversity of HIV genomes in small cohorts of patients (n < 250)[11-24, 48], our analyses evaluated HIV genome-wide diversity using 2996 full-length genomic sequences sampled from 1705 patients worldwide. We evaluated three important aspects of HIV genomic diversity using the integrated datasets of genomic sequences, protein structures, HIV-human protein interactions, human immune epitopes and HIV-derived peptide inhibitors. Firstly, we quantified HIV genomic diversity at the individual and population levels. Secondly, we reported possible associations between HIV-1 amino acid diversity and protein multimerization, immunological constraints and HIV-human protein interactions. Thirdly, we mapped conserved regions in the HIV genome and characterized experimental and clinically used HIV-derived peptide inhibitors [7].

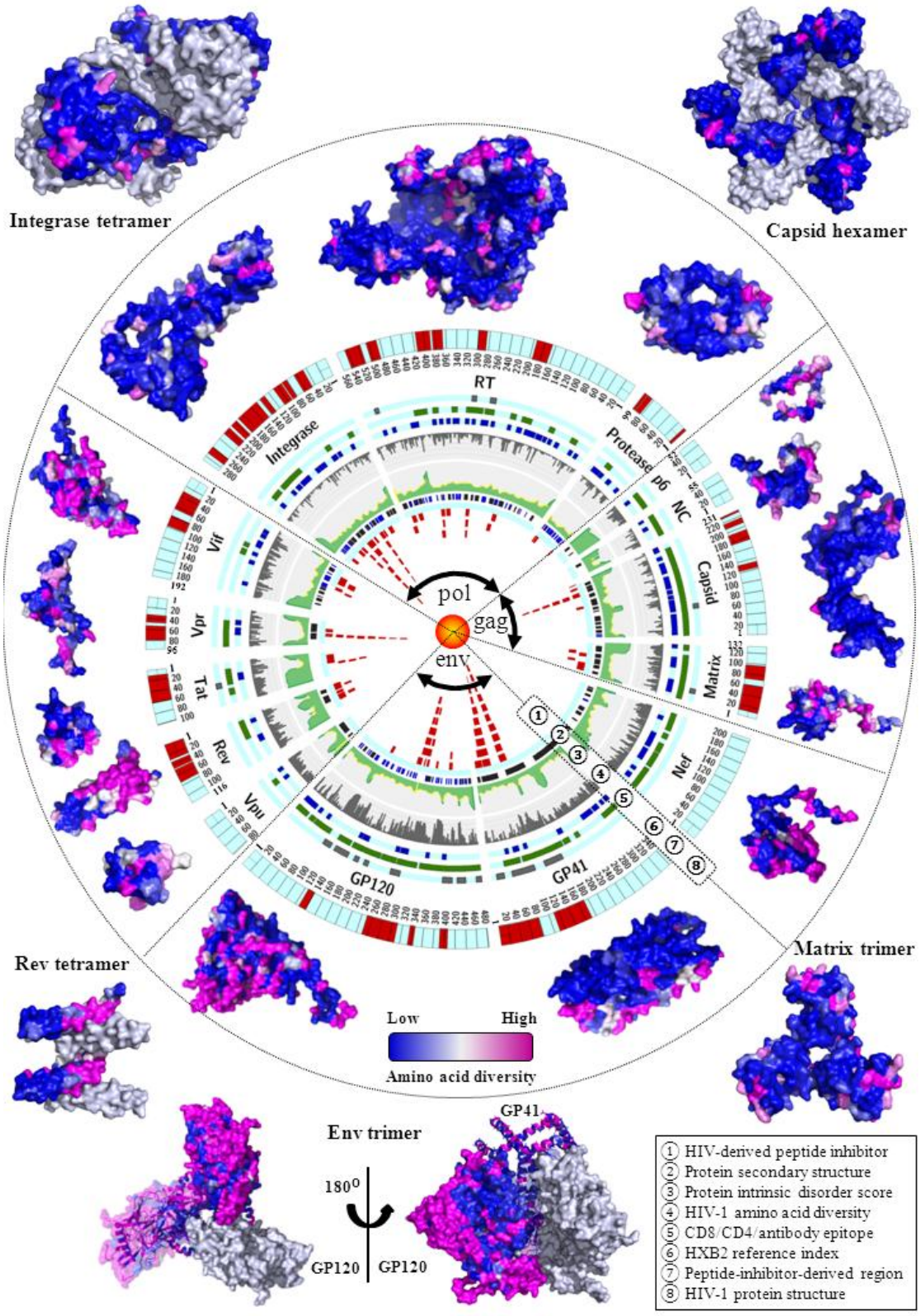


Figure 3.6: HIV-1 genomic diversity and protein structures. All 15 HIV-1 proteins are mapped in the circle with 8 layers, showing the schematic view of HIV-1 peptide inhibitors (layer 1), the protein secondary structures (layer 2, dark blue: helices, light blue: beta-strands, white: random-coil structures), protein disorder scores (layer 3), amino acid diversity of residue positions (layer 4), human CD4+/CD8+/antibody epitope regions (layer 5, three sub-layers from inside to outside represent CD8+ T cell,

CD4+ T cell and antibody epitope regions), HXB2 reference indices (layer 6), peptide-inhibitor-derived regions (layer 7) and the protein structures are colored according to the amino acid diversity of residue positions (layer 8, low: 0%, high: ≥30%). Three major genes (*gag*, *pol*, *env*) are annotated in the center. Structures of multimeric HIV-1 proteins are shown outside the circle and different protein units are colored separately. The list of PDB data is available in **Table 3.2**. Visualization: Circos V0.64 (http://circos.ca/).

Quantification of HIV genomic diversity: HIV-1 genomic diversity is the lowest within single patients and increases in the following order when different patients are considered: within subtypes, between subtypes, between groups and between HIV types (**Figure 3.1**). A nucleotide genomic diversity was quantified to be 48.3% between HIV-1 and HIV-2, 37.5% between HIV-1 groups, 14.7% between HIV-1 subtypes, 8.2% within HIV-1 subtypes, and 0.6% within single patients infected with HIV-1. These results are in good agreement with previous studies which analyzed less than 100 sequences [13, 23]. Our study quantified genomic diversity at the population level using the largest sequence dataset ever analyzed, thereby resulting in robust and accurate estimations. As shown in **Figure 3.2**, the degree of HIV genetic diversity varied along the full-length genome. A comparison of the amino acid diversity of HIV proteins revealed the highest diversity in the envelope proteins, followed by the regulatory, accessory, structural and enzymatic proteins (**Figure 3.4**, **Table 3.5**). Estimated amino acid diversities for Gag (intra-subtype: 6.6 ± 1.2%), Pol (5.7 ± 0.9%) and Env (18.7 ± 2.7%) displayed higher values than previous reports analyzing fewer than 100 sequences of subtypes A and B [24]. Using large-scale sequence datasets, our study thus provides a better estimation of genetic diversity in HIV proteins.

HIV genomic diversity is shaped by multiple factors: HIV genomic diversity is driven by the high rates of viral replication, recombination and mutation [49], but other factors also play a role in shaping HIV genomic diversity. To evaluate potential factors, we correlated HIV amino acid diversity with protein multimerization, human immunological constraints and HIV-human protein interactions. Firstly, we found that the average amino acid diversity was significantly lower in the multimeric than in the monomeric proteins, suggesting that protein multimerization places a constraint on HIV-1 sequence variability. Previous findings on other protein families have also shown that multimeric proteins are relatively conserved and have less tolerance for

amino acid substitutions [50-52]. Secondly, we showed that CD4 T cell and antibody epitope positions in the HIV-1 genome were likely to have high amino acid diversities, supporting the hypothesis that human immune system imposes a diversifying selective pressure on the HIV-1 genome [26]. Thirdly, we mapped 1352 HIV-human protein interactions between 15 HIV-1 proteins and 1052 human proteins. A strong association was found between the amino acid diversity of HIV-1 proteins and the number of HIV-human protein interactions (**Figure 3.4**). HIV-1 proteins with higher genetic diversities tended to interact with more human proteins. This is likely associated with structurally disordered regions in HIV-1 proteins (**Figure 3.4**B), which provide the structural flexibility for HIV to interact with multiple human proteins [40]. For instance, GP120 uses five hypervariable loops (**Figure S 3.13**) to interact with various human proteins [53]. An intricate landscape of HIV-human protein complexes is made by HIV to exploit human cellular machineries during the HIV infection and production [54]. Despite the high variability of HIV, it is surprising that the nucleotide and amino acid compositions were remarkably constant across all HIV-1 and HIV-2 clades (**Figure 3.3**A, 3B), suggesting that other constraints may be active to restrict the HIV genetic diversity [25].

Conserved drug targets in the HIV-1 genome: Many peptide inhibitors derived from HIV-1 proteins have shown promising antiviral activities and some of these inhibitors are currently under clinical trials [55, 56]. Our study summarized HIV-derived peptide inhibitors published between 1993 and 2013 (**Figure S 3.13-Figure S 3.22**, **Table S 3.1**), and mapped the positions of these inhibitors to the HIV-1 genome (**Figure 3.6**). We showed that most peptide inhibitors were derived from the regions of HIV-1 subtype B proteins (**Figure 3.5**B), which had conserved, solvent exposed and intrinsically ordered structures (**Figure 3.5**E). This information enhances current understanding of HIV-derived peptide inhibitors, which may provide valuable guidelines for the design of novel peptide inhibitors [57, 58]. In the full-length genome, we identified conserved regions in Capsid, Nucleocapsid, Protease, RT, Integrase, Vpr and N-terminal domain of GP41 (**Figure 3.2**). These conserved regions have been targeted by known anti-HIV inhibitors (**Figure 3.6**). For instance, over 40 experimental inhibitors with promising antiviral activities have been designed to target Capsid and Nucleocapsid [9]. HIV enzymes (Protease, RT, Integrase) are targeted by most of the FDA-approved antiretroviral drugs. Peptide inhibitor T20

targets the N-terminal heptad domain of GP41 [59]. Overall, our sequence analysis mapped the conserved drug target regions in the HIV-1 genome, providing useful information for drug design.

Implications for HIV vaccine development: HIV subtype- and geography-specific vaccination has been proposed to contend with the challenges imposed by the high HIV genetic diversity [2]. Previous vaccine trials were carried out in regional populations dominated by a single HIV-1 subtype or CRF. For instance, the STEP [11] and RV144 [12] vaccine trials targeted patient populations mainly infected by subtype B and CRF01_AE, respectively. Particularly, the RV144 trial in 2009 showed the first sign that a prime–boost strategy achieved a modest vaccine efficacy (31.2%) in the heterosexual population, which was at risk for infections with CRF01_AE [12, 48]. In our analysis, CRF01_AE has the lowest genomic diversity among the 12 analyzed HIV groups, subtypes and CRFs (**Figure 3.1D, Figure 3.2A, 2B**). It is thus tempting to speculate that the low diversity of CRF01_AE may have contributed to the success of the RV144 trial. As conserved epitopes are ideal targets for potential vaccines to contend with the high HIV diversity [24, 60], our study highlighted position-specific conservation along the full-length HIV genome (**Figure 3.6**). Moreover, HIV-1 consensus sequences have been considered as potential vaccine candidates to minimize genetic diversity between vaccine candidates and circulating strains [2]. Previous analyses on fewer than 100 Matrix and GP160 sequences reported that genetic diversity between subtype-specific consensus sequences and circulating strains was only half of the genetic diversity between circulating strains from the same subtype [2]. We found that in the full-length HIV genome, this effect was much smaller as we only observed a 32.5% reduction of the genomic diversity (8.3% vs. 12.3%, **Figure S 3.23**). As the most explored vaccine target protein, GP120 has the highest genetic diversity among all HIV proteins (**Table 3.5**), presenting a challenge in the search for broadly neutralizing antibodies and vaccines [61]. Furthermore, we mapped the global distribution of HIV-1 genomic diversity (**Figure S 3.5**). Our results showed the highest HIV genomic diversity in Central Africa, the birthplace of HIV [1, 29], which suggested the difficulty of implementing HIV vaccines in this region.

Limitations and future perspectives: The limited number of genomic sequences in HIV-1 subtypes H, J and K, group P and HIV-2 group B ($n<10$) may have affected

our estimation of sequence diversity, but consistent patterns were detected in the full-length genome across different HIV groups and subtypes (**Figure 3.2**). Our structural analysis focuses on HIV-1 proteins because most PDB data is available for HIV-1 but not for HIV-2. Information on positions involved in HIV-human protein interactions is largely lacking, restricting our analysis from exploring the genetic diversity of protein interaction positions. Beside the multiple factors described in our study, other driving forces may shape HIV genetic diversity [25] and the genetic diversity data reported in our study can be useful for further investigations. Despite an extensive search, anti-HIV peptide inhibitors other than the ones described here may have been developed, but major changes in our conclusions regarding the known peptide-derived regions are not expected. Future studies are still needed to clarify how to improve vaccines and anti-HIV inhibitors based on the information of HIV genomic diversity. The increased knowledge of genome-wide diversity from our study may contribute to a better rational design of HIV vaccines and inhibitors.

3.6 Additional file 1: Figures

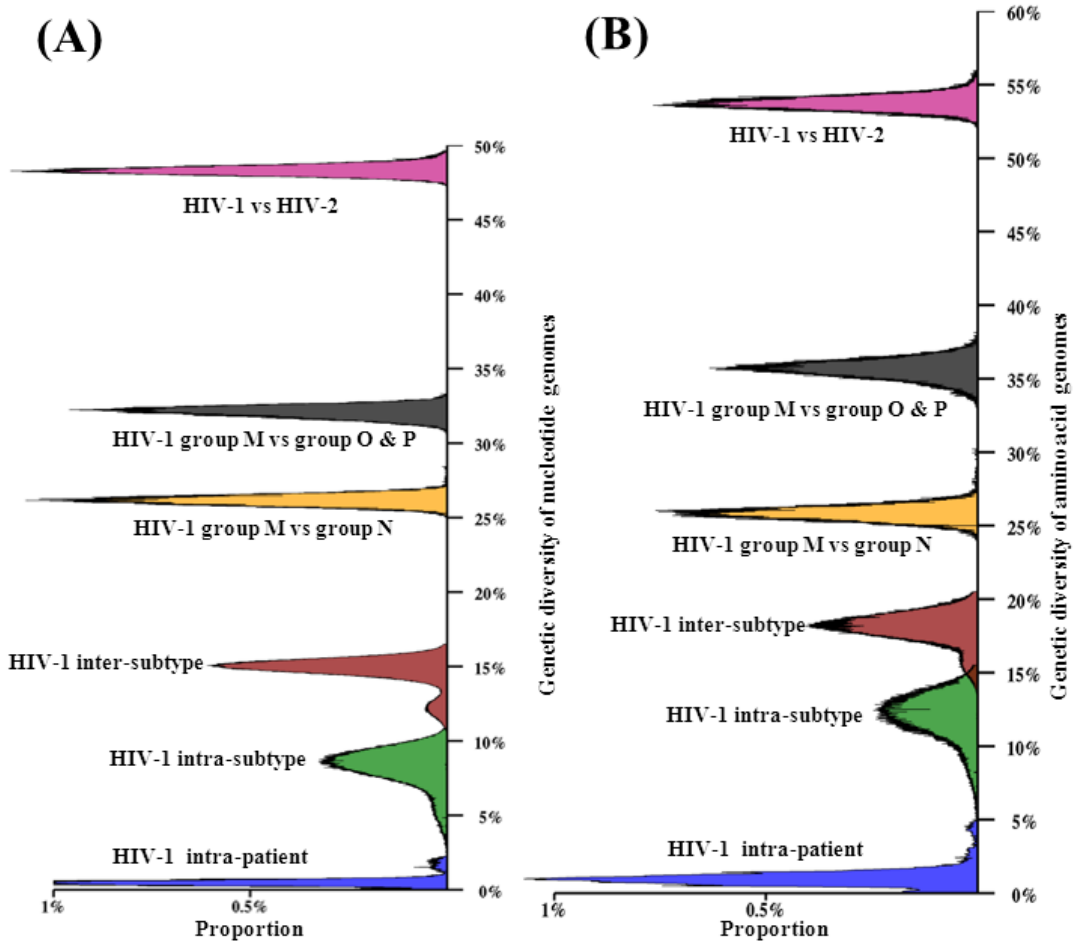


Figure S 3.1: Distribution plots of nucleotide (A) and amino acid (B) genomic diversity among HIV types, groups and subtypes. See figure captions in **Figure 3.1**. Mean value of genomic nucleotide diversity: 48.32% (CI: 47.80-48.89%) between HIV-1 and HIV-2, 37.48% (CI: 25.98-45.70%) between HIV-1 groups, 14.72% (CI: 12.19-15.79%) between HIV-1 subtypes, 8.20% (CI: 5.32-9.95%) within HIV-1 subtypes and 0.6% (CI: 0.2-1.4%) within HIV-1 patients. Mean value of amino acid genomic diversity: 53.75% (95% CI: 52.97-54.57%) between HIV-1 and HIV-2, 41.11% (CI: 25.58-54.28%) between HIV-1 groups, 18.02% (CI: 15.59-19.60%) between HIV-1 subtypes, 11.99% (CI: 8.63-14.36%) within HIV-1 subtypes and 1.1% (CI: 0.3-2.2%) within HIV-1 patients. For visualization purpose, we did not plot the long peak region of HIV-1 intra-patient nucleotide diversity which is above 1% in the proportion of genomic diversity.

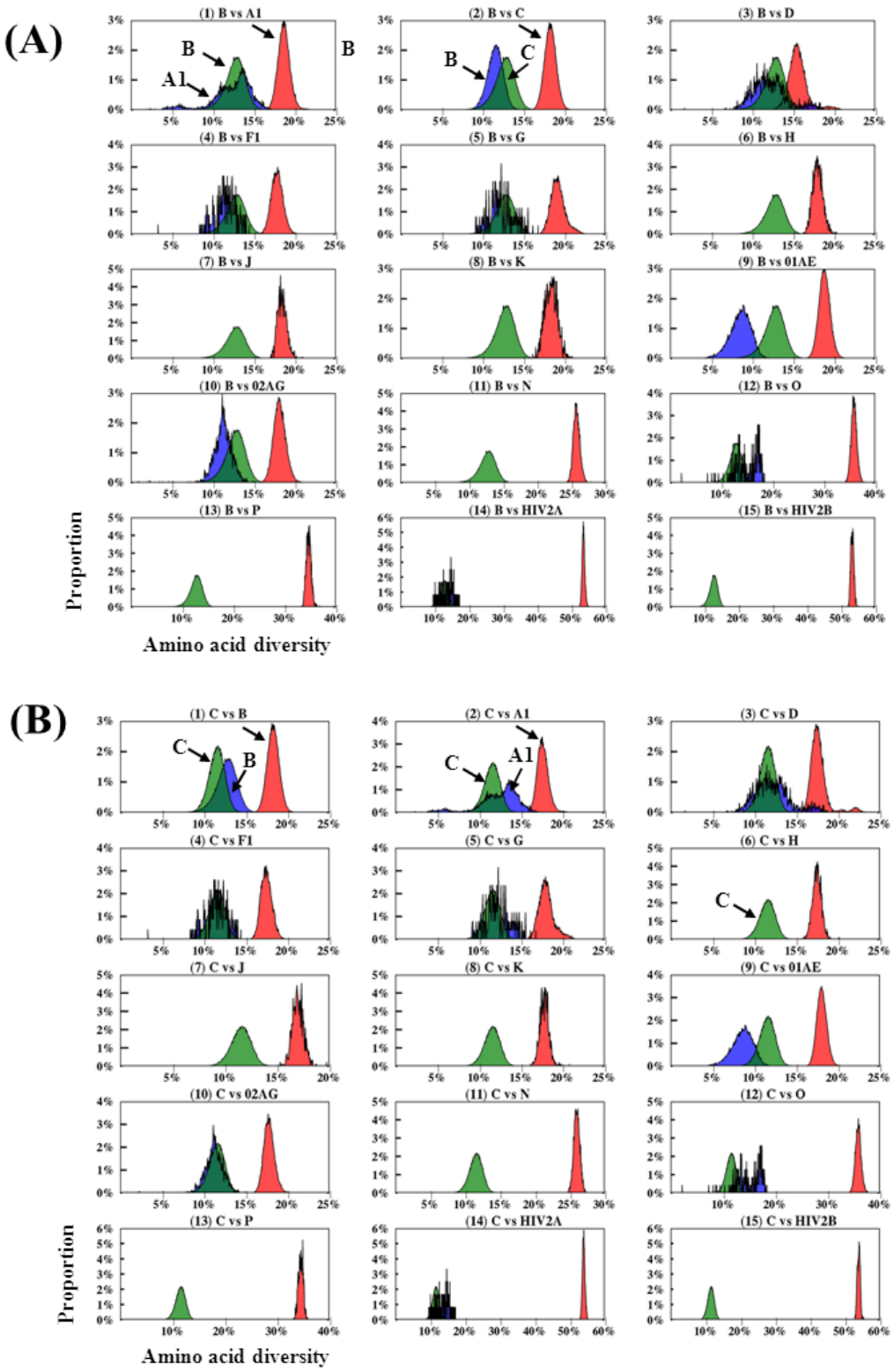


Figure S 3.2: (A) Distribution plots of amino acid diversity between HIV-1 subtype B and other clades (HIV-1 subtypes, HIV-1 and HIV-2 groups) in the protein coding regions of the full-length genome. Each graph displays the distribution plots of diversity within subtype B (green), diversity within the other subtype/group (blue) and diversity between subtype B and the other subtypes/groups (red). The x- and y-axes indicate the HIV amino acid diversity and the proportions, respectively. (B)

Distribution plots of amino acid diversity between subtype C and other clades (HIV-1 subtypes, HIV-1 and HIV-2 groups).

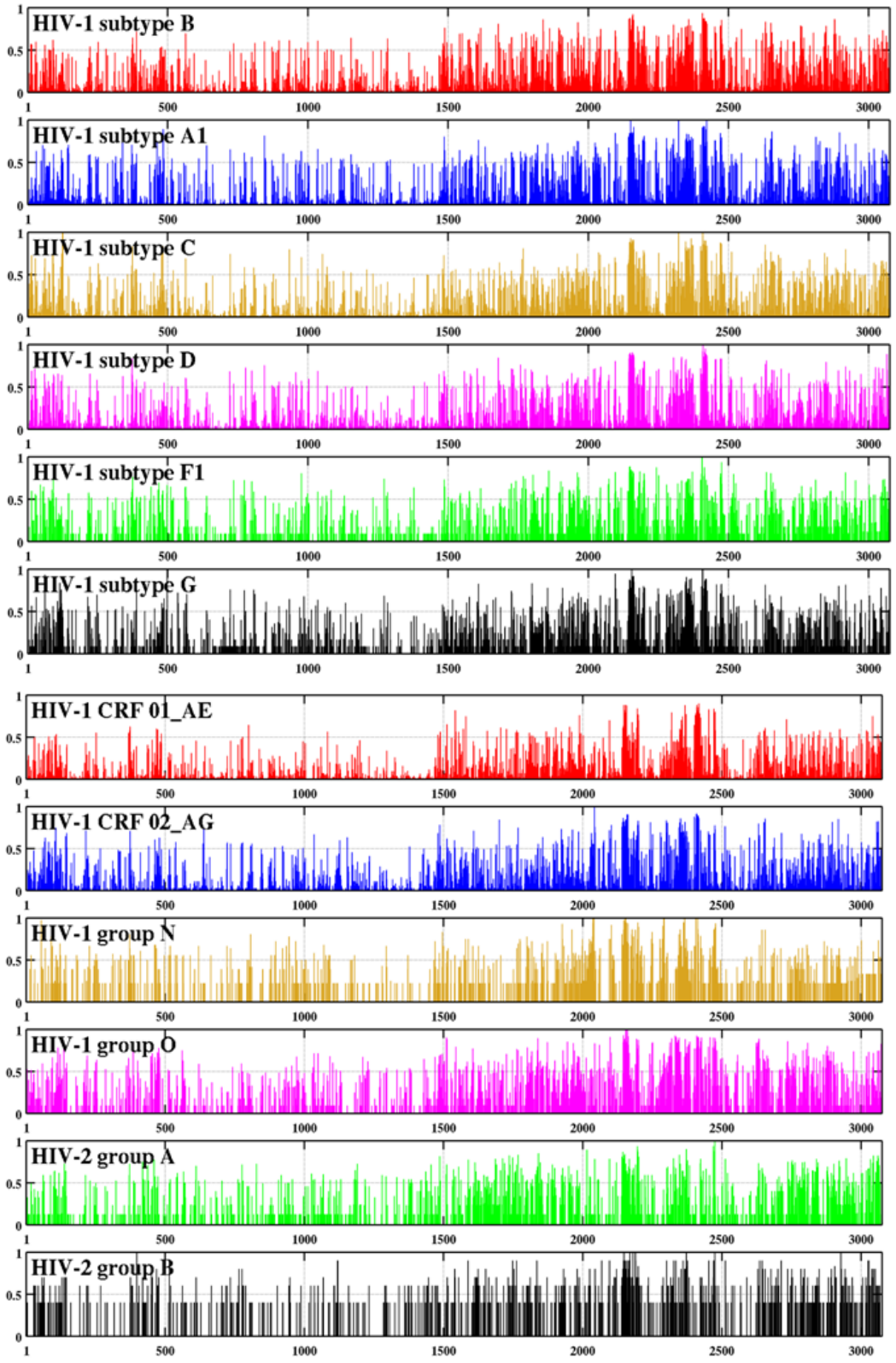


Figure S 3.3: Amino acid diversity along the full-length HIV genome. Twelve subplots

individually show the nucleotide diversity results for subtype B, A1, C, D, F1, G, CRF01_AE, CRF02_AG, and HIV-2 group A and B, HIV-1 group N, O.

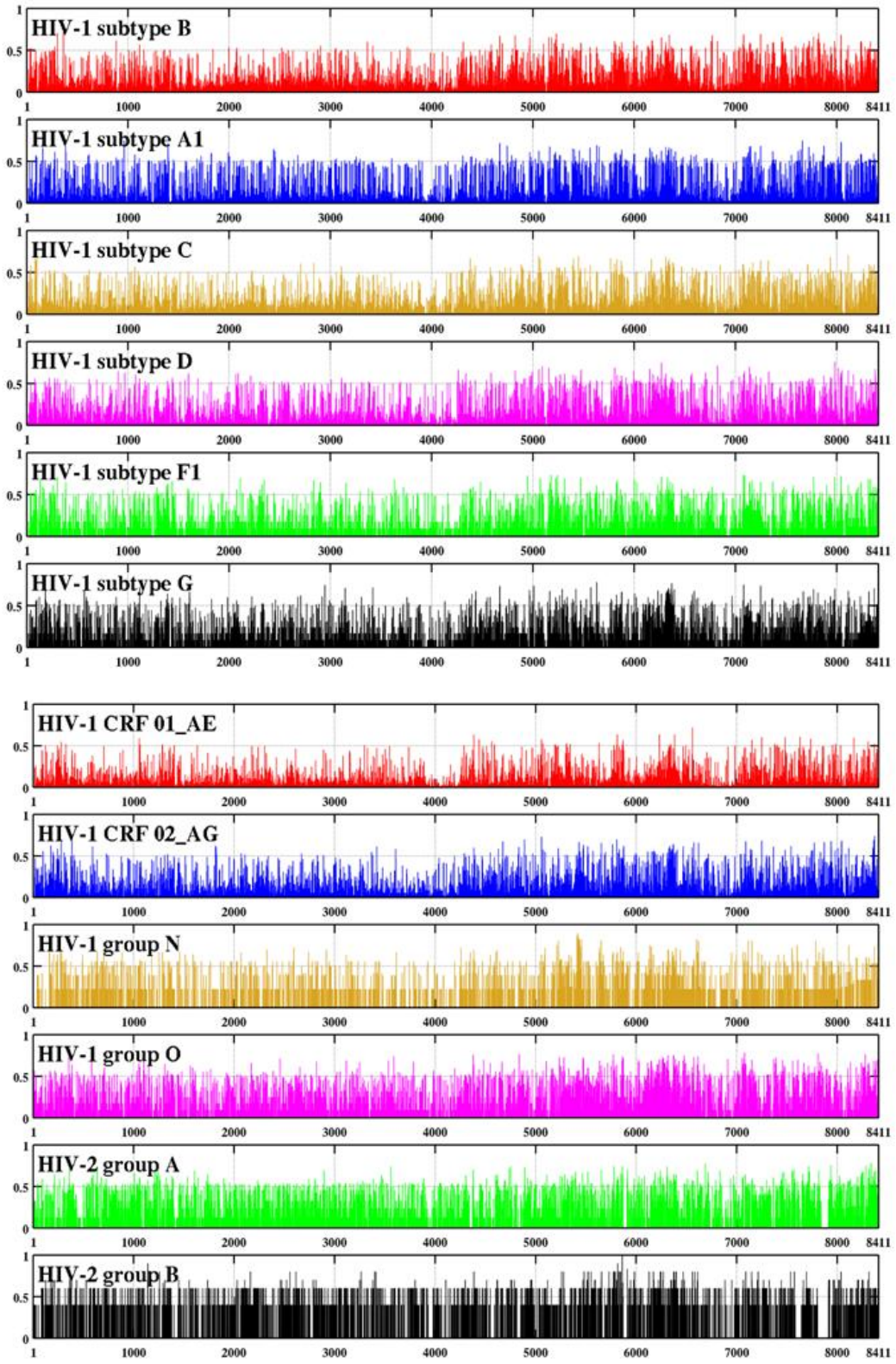


Figure S 3.4. Nucleotide diversity along the full-length HIV genome. Twelve subplots individually show the nucleotide diversity results for subtype B, A1, C, D, F1, G, CRF01_AE, CRF02_AG, and HIV-2 group A and B, HIV-1 group N, O.

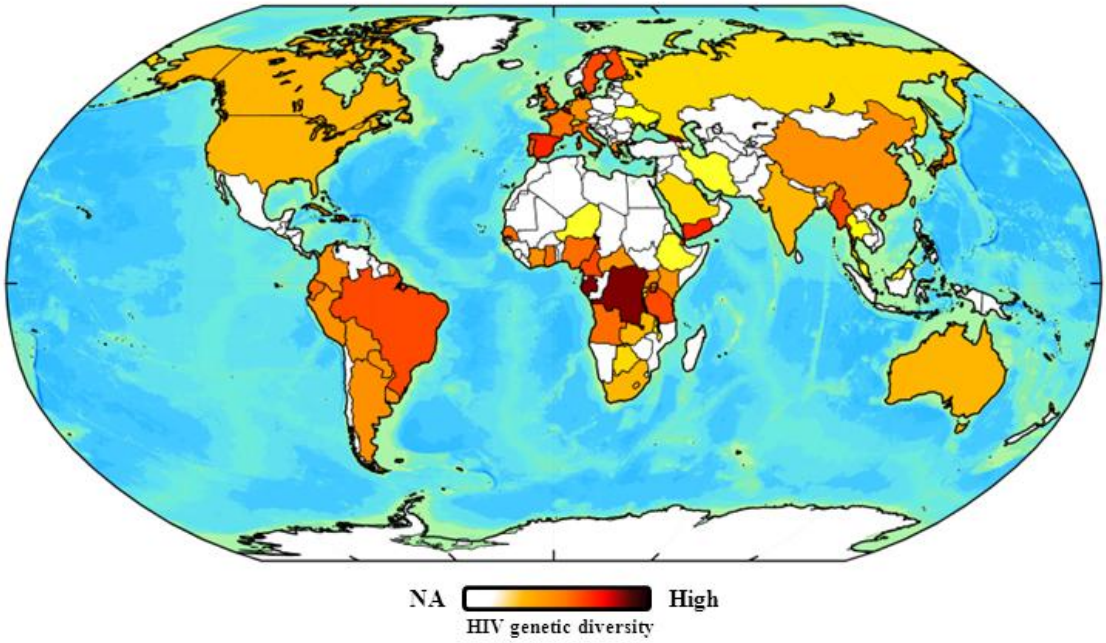


Figure S 3.5: Global distribution of HIV-1 genomic diversity. Countries with no sequences available (NA) are colored white. Amino acid genomic diversity in individual countries was mapped onto the global cartographic map in Natural Earth V2.0.0 (http://www.naturalearthdata.com/). Countries with infections by different groups or subtypes had higher genomic diversity, with the highest being found in Central Africa. Our results are consistent with the known distribution of HIV-1 subtypes described in [29], implying that the strains included in our study may capture the global HIV-1 diversity.

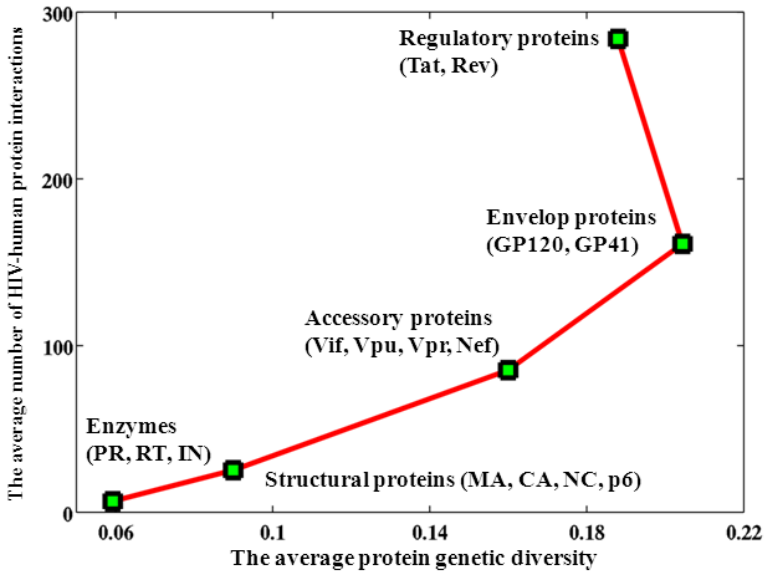


Figure S 3.6: Average amino acid diversity of HIV-1 protein clusters and number of HIV-human protein interactions. The five protein clusters include: viral enzymes (PR, RT, IN), accessory proteins (Vif, Vpu, Vpr, Nef), envelope proteins (GP120, GP41) and regulatory proteins (Tat, Rev). Proteins are clustered according to their functional roles in the HIV-1 life cycle [4].

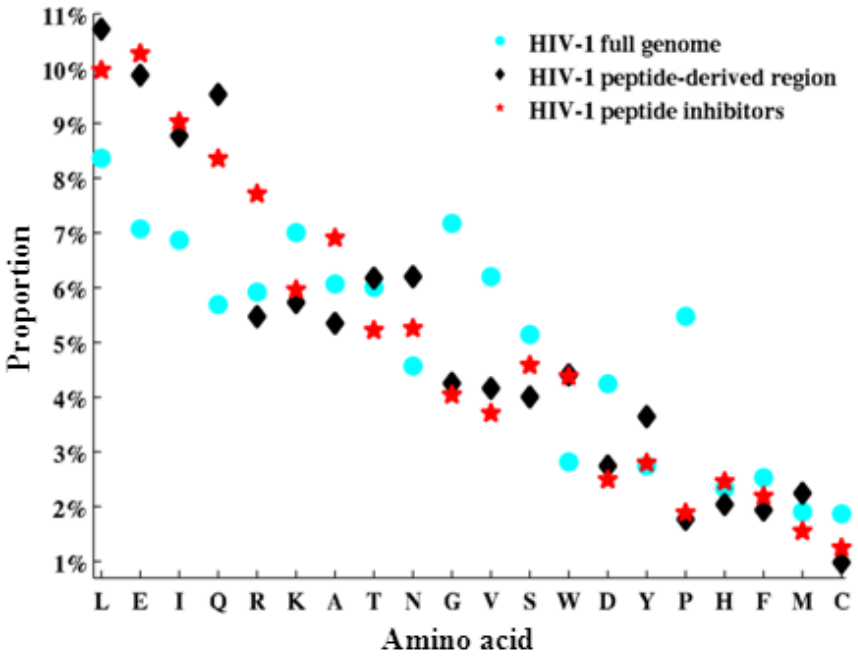


Figure S 3.7: Amino acid composition of HIV-1 subtype B genome (blue circles), HIV-1 peptide-derived regions (black diamonds) and the HIV-1 peptide inhibitor sequences (red stars).

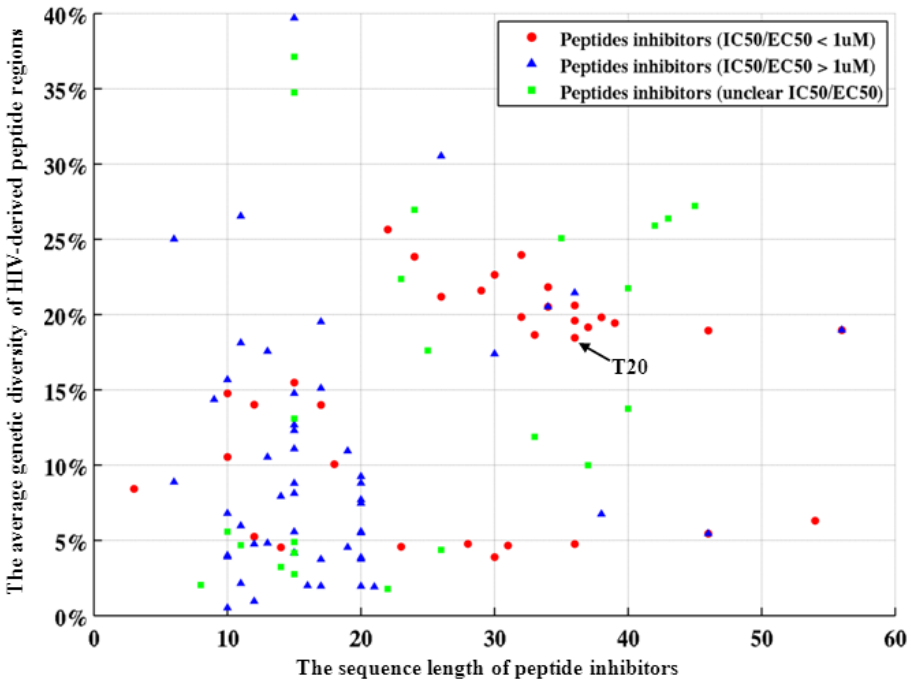


Figure S 3.8: Average genetic diversity of peptide-derived regions in HIV-1 subtype B. The x-axis indicates the length of peptide inhibitor sequences. The y-axis indicates the average genetic diversity of the known peptide-derived regions of HIV-1 subtype B genome.

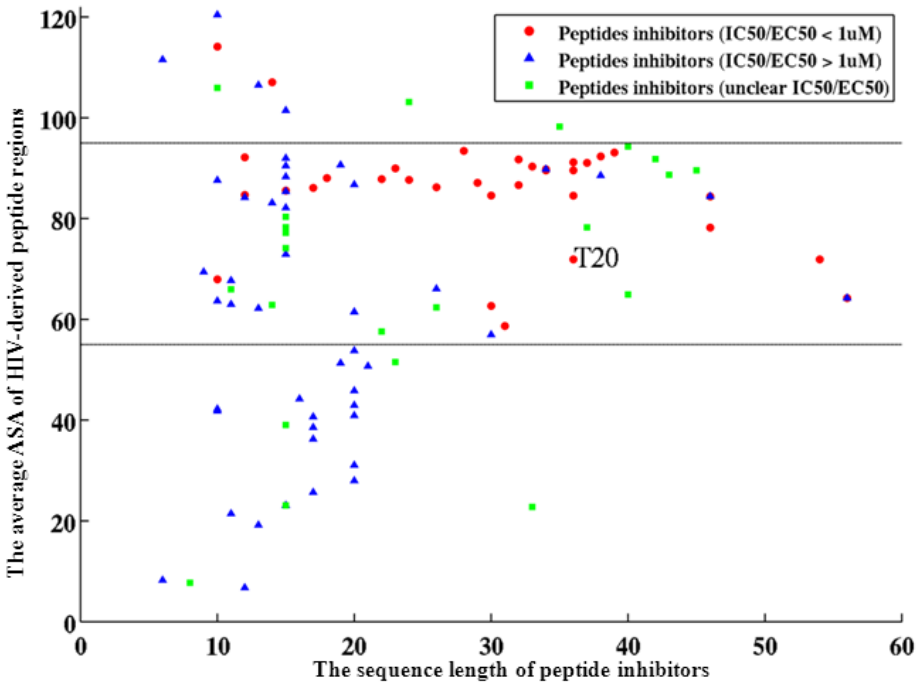


Figure S 3.9: Solvent accessible surface area (ASA) of peptide-derived regions in HIV-1 subtype B. The x-axis shows the length of peptide inhibitor sequences. The y-axis shows the average ASAs (Å) of known peptide-derived regions of HIV-1 subtype B genome. Horizontal lines mark the average ASAs of 55 Å and 95 Å, covering most peptide inhibitors with IC50/EC50<1μM.

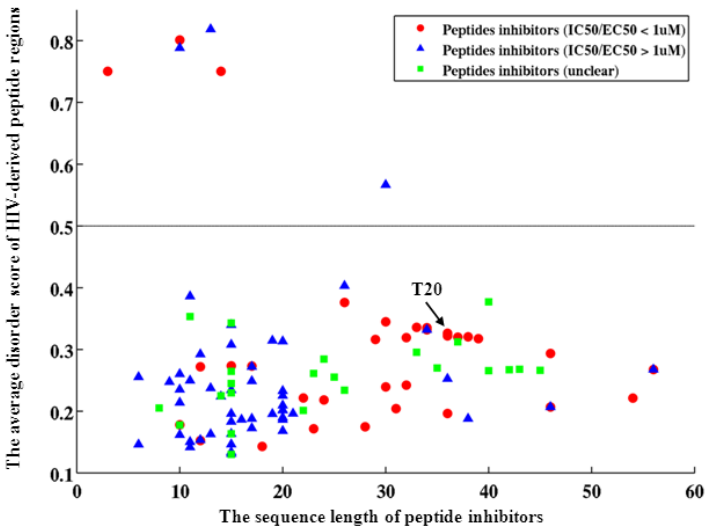


Figure S 3.10: Protein intrinsic disorder scores of peptide-derived regions in HIV-1 subtype B. X-axis indicates the length of peptide inhibitor sequences. Y-axis indicates the average protein intrinsic disorder scores in the known peptide-derived regions of HIV-1 subtype B genome. The horizontal line at the value of 0.5 indicates the cutoff of the disorder score for determining disordered (≥0.5) or ordered (<0.5) structural regions.

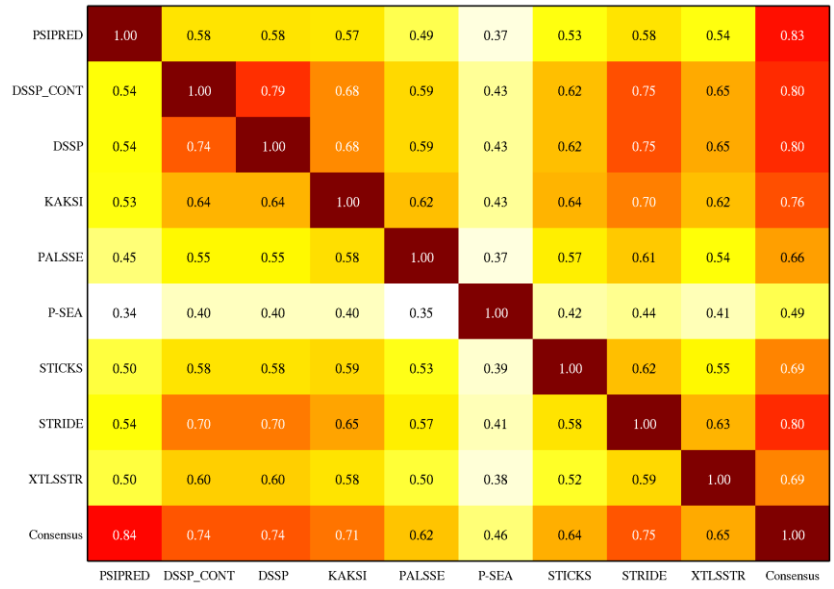


Figure S 3.11: Similarity of prediction results between the consensus and the 9 protein secondary structure methods. Consensus assignments were obtained using the majority voting strategy among the 9 individual methods. Given 15 HIV-1 proteins in the full-length genome of HIV-1 subtype B, similarities between two methods were calculated by the percentages of common predictions of alpha-helix (top-right part of matrix) and beta-strand (left-bottom part of matrix) structures.

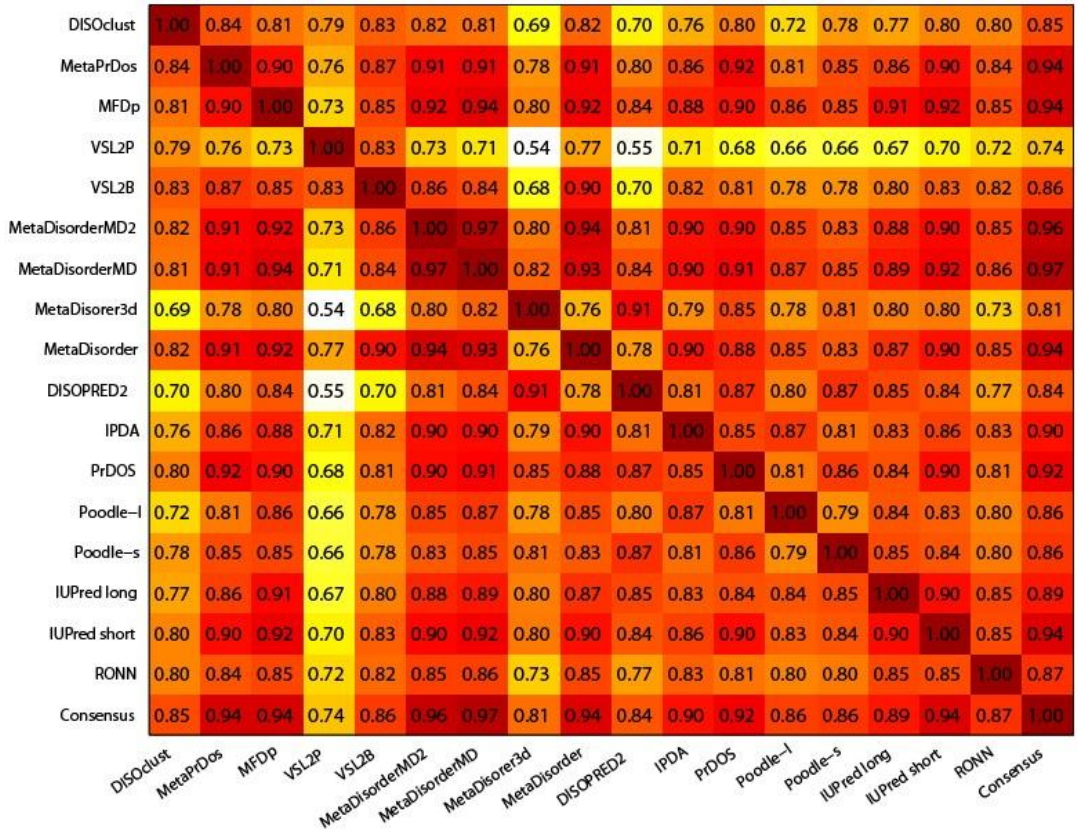


Figure S 3.12: Prediction similarities of the consensus and 17 methods for protein intrinsically disorder prediction. Prediction similarities were calculated by

the percentages of common predictions of ordered (disorder tendency score < 0.5) or disordered (disorder tendency score ≥ 0.5) positions in HIV-1 protein structures. Consensus predictions were obtained using the majority voting strategy among the 17 individual methods. The consensus method has the highest average prediction similarities compared to the other methods.

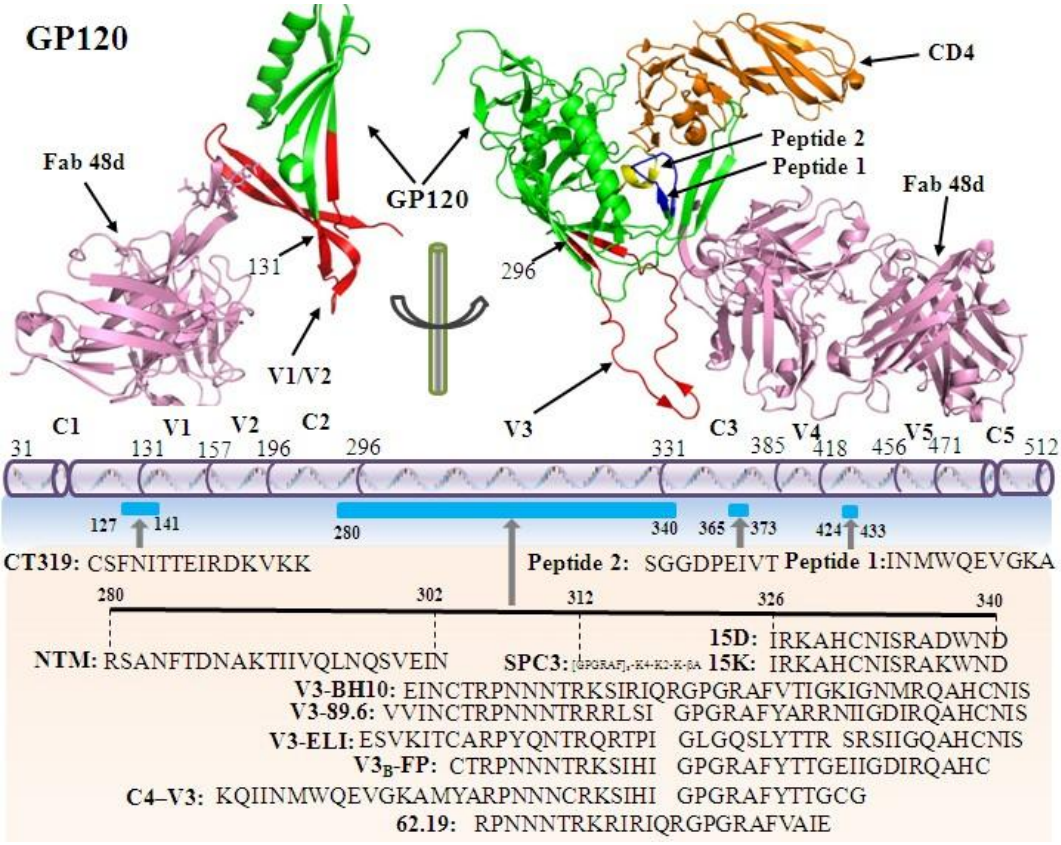


Figure S 3.13: Structure of the HIV-1 GP120-CD4-Fab 48d complex (PDB: 2B4C, 3U4E) and mapped GP120 peptide-derived inhibitors. On the structure, CD4 and Fab 48d structures are colored orange and pink, respectively. The GP120 and peptide inhibitor sequences is annotated beneath the protein structures. Peptide inhibitors are mapped to the GP120 functional domains (bottom), including 5 variable domains (V1-V5) and 5 conserved domains (C1-C5) [62].

The V1 to V3 and V5 loops have been identified as the minimal functional units of GP120 to mediate CXCR4-dependent infection [62]. The V3 loop is the major target for neutralizing antibodies and V3-derived peptides offer promising anti-HIV activities [63]. GP120-derived peptides can inhibit the interactions between GP120 and T-cell surface glycoproteins (e.g. CD4, CD19), chemokine co-receptors (e.g. CCR5, CXCR4) and monoclonal antibodies [63]. The inhibition activity of GP120-derived peptides can be strain-dependent and cell-dependent (**Table S 3.1**).

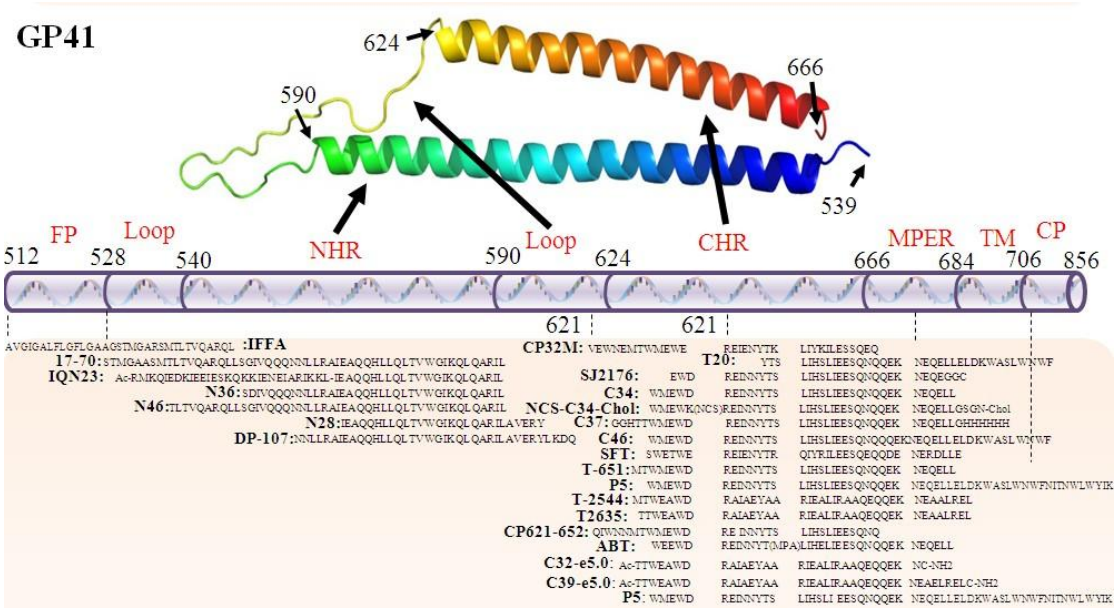


Figure S 3.14: GP41 structure and mapped GP41-derived peptide inhibitors. The protein cartoon representation of GP41 protein is shown on top and the peptide inhibitor sequences are annotated beneath the structures. SIV GP41 structure is visualized (PDB: 2EZS) since the crystal structure of HIV-1 GP41 has not yet been fully resolved (February 2014).

GP41 forms a trimer with GP120 in a viral spike on the surface of HIV particles (**Figure 3.5**). Three C-terminal heptad repeats (CHRs) bind with three N-terminal heptad repeats (NHRs) to form a 6-helix bundle, which fuses the viral with cellular membrane to create a fusion pore during viral entry [64, 65]. Peptides derived from either NHR or CHR can mimic the viral structures to prevent viral entry. T20 (Enfuvirtide, Fuzeon, DP178), derived from NHR of HIV-1 subtype B strain LAI, is an 36AA L-peptide inhibitor which targets multiple sites on GP41 and GP120 [59]. Most GP41 peptide inhibitors have been derived from the conserved pre-hairpin structure [66]. GP41-derived L-peptides have many defects such as proteolytic degradation, limited potency and toxicity [67]. To overcome the defects of L-peptides, D-peptides have been proposed as promising fusion inhibitors (reviews [68, 69]). For instance, the D-peptide inhibitors PIE12 [70] and IQN17 [71] have shown good anti-HIV activities by mimicking the heptad repeat regions in GP41.

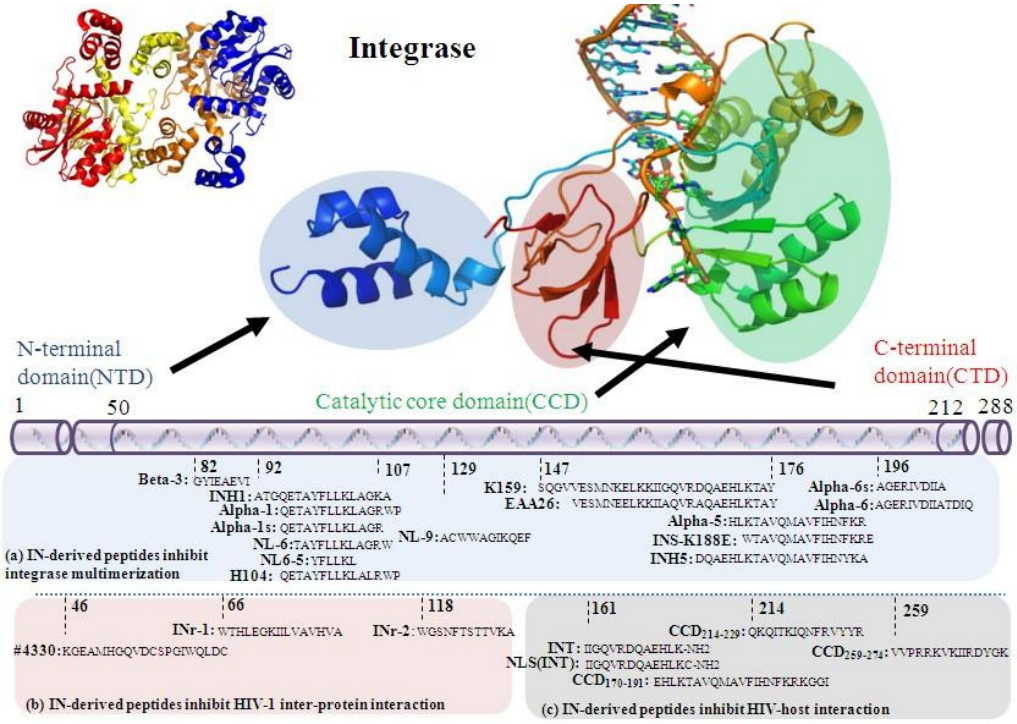


Figure S 3.15: HIV-1 Integrase tetramer (PDB: 1K6Y) and mapped integrase-derived peptide inhibitors. The protein cartoon representation of Integrase domains is shown on top and the peptide inhibitor sequences are annotated beneath the structures. For visualization purpose, we used the integrase monomer from the prototype foamy virus (PDB: 3OY9). The HIV-1 integrase structure has an N-terminal domain (NTD), a catalytic core domain (CCD) and a C-terminal domain (CTD), which are connected by flexible links [4]. Integrase plays multiple roles during viral reverse transcription and integration [72]. The key role of integrase is to insert viral dsDNA into human chromatins, creating a viral reservoir for viral infection [73]. Integrase-derived peptides can inhibit integrase-mediated catalytic functions, integrase inter-domain interactions and/or integrase-human protein interactions (**Table S 3.1**). Peptide inhibitors derived from CCD can inhibit integrase dimerization, 3'-end DNA processing and strand transfer during viral integration [74].

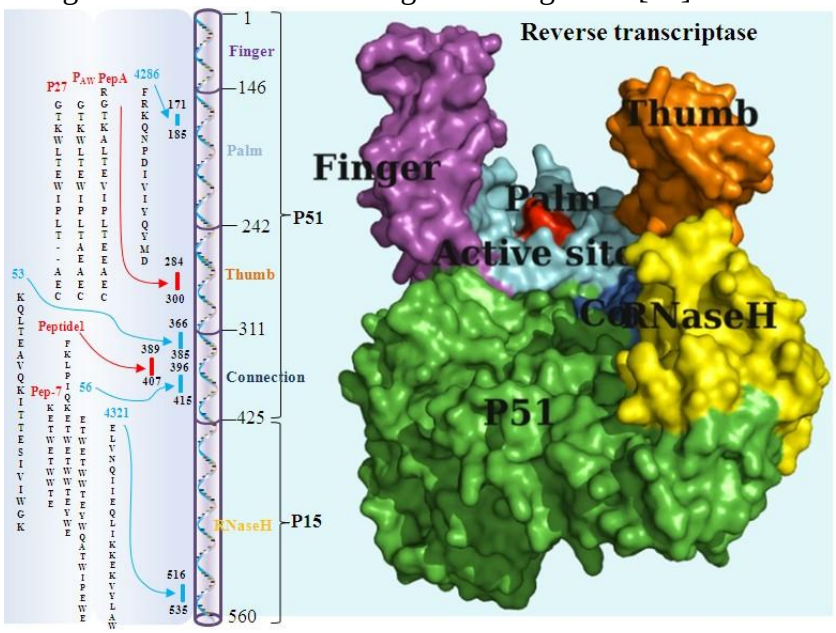


Figure S 3.16: HIV-1 reverse transcriptase (RT) structure (PDB: 3DLK) and mapped RT-derived peptide inhibitors. The surface representation of reverse transcriptase domains is shown on right and the peptide inhibitor sequences are annotated on the left side of the structure.Reverse transcriptase forms a heterodimer to synthesize dsDNA from the viral genomic RNA [75]. RT structures are comprised of the finger, palm, thumb, connection, RNaseH and P51 functional domains [76]. Peptide inhibitors derived from the connection (Pep-7 [77], Peptide1 [78]) and thumb domain (P24[79], P27[79], P_{AW}[79]) can block the dimerization of p66 and p51. Nanoparticle delivery systems can improve the delivery of RT peptide inhibitors [79].

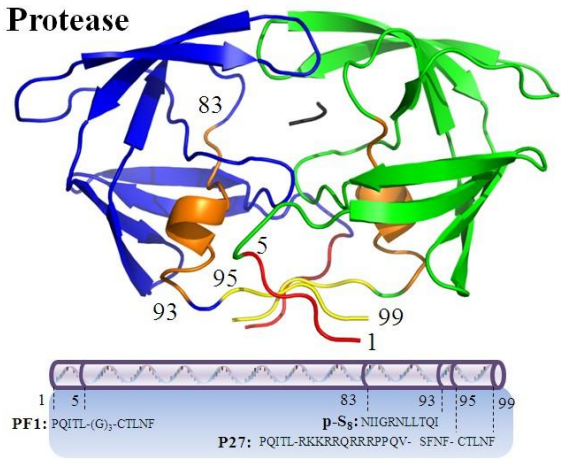


Figure S 3.17: HIV-1 Protease homodimer structure (PDB:1A30) and mapped protease-derived peptide inhibitors. The protein cartoon representation of protease is shown on top, and the peptide inhibitor sequences are annotated beneath the structure. Beta-sheets of the N-terminal and C-terminal domains are crucial for protease dimerization [80, 81]. HIV-derived peptide inhibitors that mimic the N- and C-terminal domains have been investigated as potential protease inhibitors. These include the cross linked interfacial peptide PF1 [82] and the PR-derived peptide p-S8 [83]. Peptides derived from protease 83-93 can inhibit protease folding [84-86].

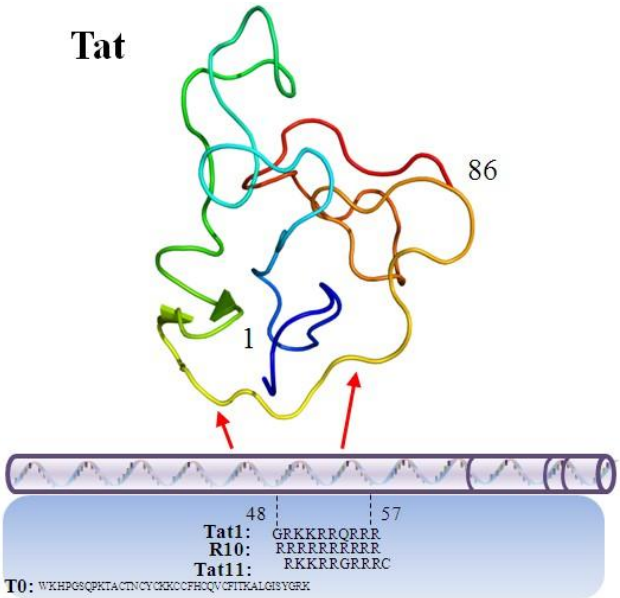


Figure S 3.18: HIV-1 Tat structure (PDB: 1JFW) and mapped Tat-derived

peptide inhibitors. The protein cartoon representation of Tat is shown on top. The peptide inhibitor sequences are annotated beneath the structure.

The regulatory protein Tat can bind with GP120 to enhance viral entry [87]. Peptide sequences derived from the Tat positions 48-57 can interrupt the Tat-GP120 interaction in a concentration-dependent manner [87]. The peptide inhibitor Tat11 can interrupt nuclear import by interacting with the host importin beta protein [88]. Moreover, the Tat-mediated transcription can be blocked by the inhibition of Tat-TAR interactions, which involves the arginine rich motif of Tat and the 3-nt bulge of the TAR RNA hairpin (U23, A27, U38) [89].

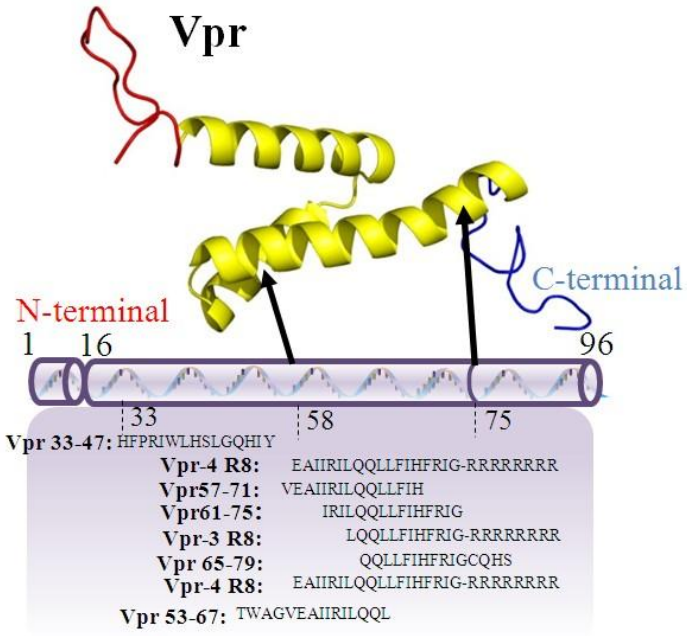


Figure S 3.19: HIV-1 Vpr structure (PDB:1M8L) and mapped Vpr-derived peptides. The protein cartoon representation of Vpr is shown on top and the peptide inhibitor sequences are annotated beneath the Vpr structure.

An interaction between Vpr and RT has not been reported, nor an interaction between Vpr and Integrase. However, peptide inhibitors derived from Vpr domains (positions: 57-71, 61-75) can interfere with the activity of both RT and Integrase [90]. Two studies have independently shown that Vpr-derived peptides (positions: 55-69, 60-74) can inhibit the strand transfer and the 3'-end-processing reactions conducted by Integrase [91, 92].

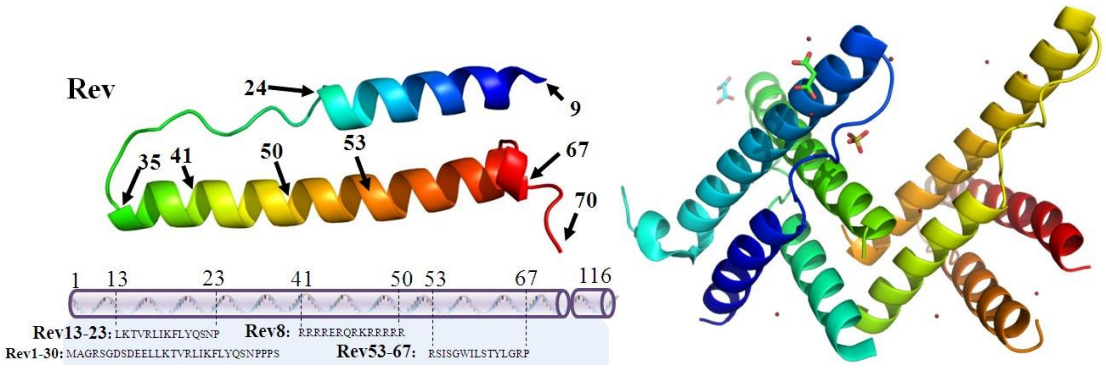


Figure S 3.20: HIV-1 Rev tetramer structure (PDB: 3LPH) and mapped Rev-derived peptide inhibitors. The protein cartoon representation of Rev is shown on top and the peptide inhibitor sequences are annotated beneath the Rev structure.

Rev can target the Rev response element (RRE) in the viral RNA genome during nuclear export, while Rev-derived peptides can interrupt the Rev-RRE interaction [93]. Rev can physically bind with Integrase to form a pre-integration complex so that viral integration can be postponed until the completion of nucleocytoplasmic shuttling [94, 95]. Two Rev-derived peptides (positions: 1-30, 49-74) can inhibit the Integrase 3'-end processing and the strand-transfer in cell-free assays [94]. Moreover, direct interactions between two Rev domains (positions: 12-23, 53-67) and integrase domains (positions: 118-128, 66-80) have been reported [96]. Two shorter Rev peptides (positions 13-23, 53-67) have also shown the inhibitory activity [97]. The Integrase-derived peptides INr-1 and INr-2 can stimulate viral genome integration and interrupt the Rev-Integrase protein interaction [98].

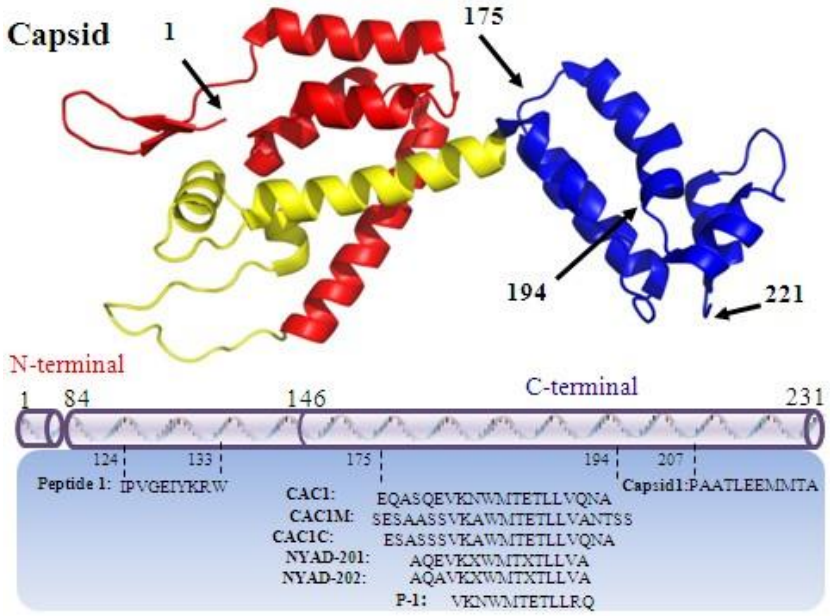


Figure S 3.21: Monomer structure of HIV-1 Capsid (PDB: 2NTE) and mapped Capsid-derived peptide inhibitors. The protein cartoon representation of Capsid is shown on top and the peptide inhibitor sequences are annotated beneath the structure.
Capsid pentamers and hexamers constitute the internal shell of viral particles [35]. The alpha-helical structure of the C-terminal domain (CTD, positions: 146-231) participates in the capsid multimerization [99]. Peptide inhibitors derived from the CTD can interrupt the multimerization of HIV-1 Capsid by mimicking the capsid multimerization interfaces. The peptide inhibitor CAC1 derived from the CTD domain can disassociate CTD dimers (K_d = 50 uM) [100]. Since peptide inhibitors must penetrate the viral membrane to prevent the Capsid multimerization, cell-penetrating peptides have been designed to improve the peptide potency in cell culture experiments [101, 102]. For instance, the peptide inhibitor CAI [103] has been converted into a cell-penetrating peptide NYAD-1, which improves the binding affinity and inhibits the post entry stage [101]. The cell-penetrating peptides NYAD-201 and NYAD-202 have shown promising anti-HIV activities [102].

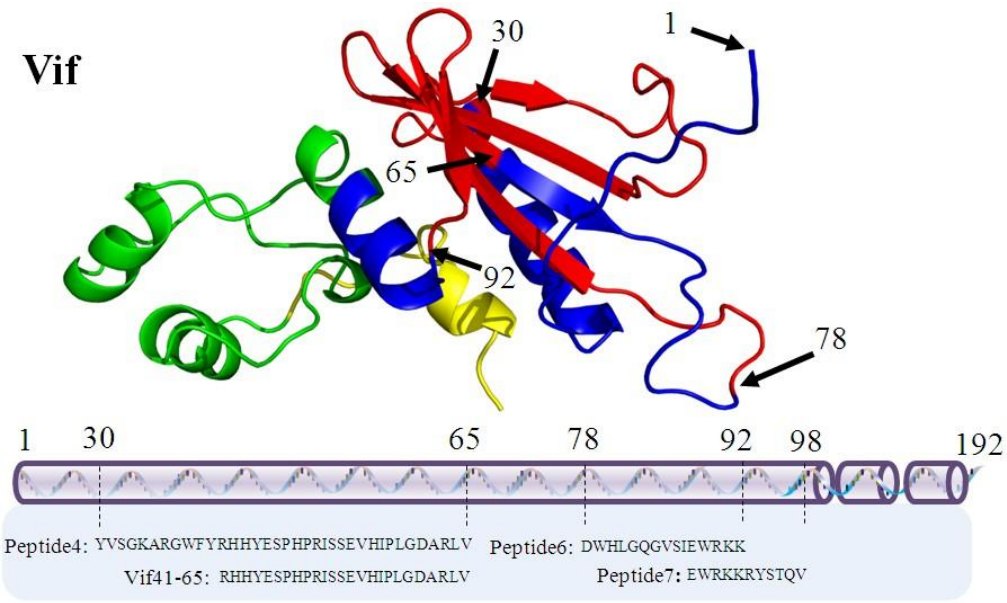


Figure S 3.22: HIV-1 Vif structure (PDB: 4N9F) and mapped Vif-derived peptide inhibitors. Peptide inhibitor sequences are annotated beneath the Vif structure. The N-terminal domain, the central domain and the C-terminal domain of Vif are colored blue, green and yellow, respectively. The Vif-derived peptide Vif41-65 (positions: 41-65) can inhibit protease activity [104]. The Vif positions (36, 47, 101, 117, 124) are associated with PI treatment [105]. Two Vif-derived peptides 30-65 and 78-98 have also been shown to inhibit the protease activity [106]. The N terminus of Protease (positions: 1-9) interacts with the central domain of Vif (positions: 78-98) [107]. Two Vif-derived peptides (positions: 81-88, 88-98) inhibit the protease activity [106, 108].

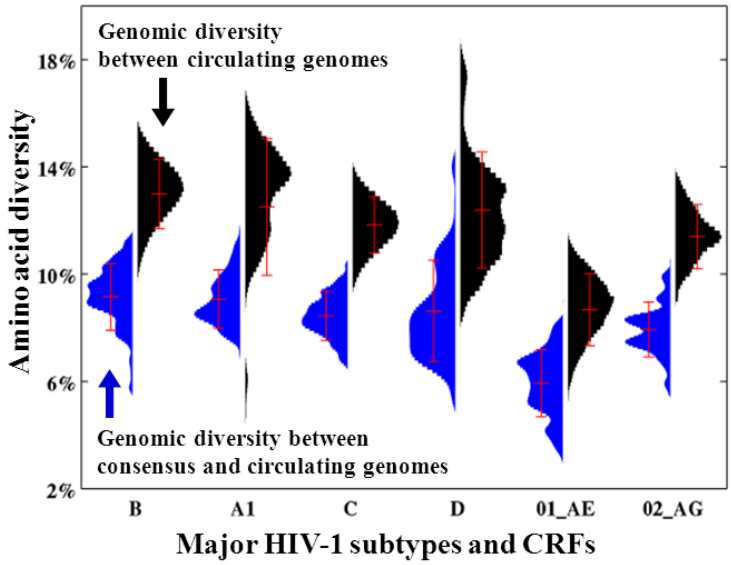


Figure S 3.23: Distribution plots of amino acid diversity between the consensus and the circulating genomes (blue), and within circulating genomes (black). The x-axis indicates HIV-1 subtypes B, A1, C, D, CRF01_AE and CRF02_AG, each of which contains more than 50 sequences in our datasets. The y-axis indicates amino acid genomic diversity. For each subtype, the consensus sequence is obtained by retaining the most prevalent residue at each position. For each of the 6 HIV-1 subtypes (A1, B, C, D, 01_AE, 02_AG), the average amino acid diversity between

circulating strains (12.3±1.5%) was significantly higher than that between the consensus and the circulating strains (8.3±1.3%, P-value<0.001).

3.7 Additional file 2: Tables

Table S 3.1: Summary of 121 peptide inhibitors derived from HIV-1 proteins. (1) Peptide position: HIV-1 proteins and HXB2 positions from which the peptide inhibitors were derived. The numbering of the GP41 peptide positions refers to the Env protein; (2) Peptide name: we used the peptide names as indicated in the published articles; (3) Sequence: peptide amino acid sequence, with additional molecule functional groups where relevant; (4) Target: target protein with which peptide inhibitors bind; (5) IC50/EC50/Kd: experimental outcomes of peptide performance in Kd, IC50 or EC50. The superscript 'd' indicates Kd and 'E' indicates EC50. Others without superscripts are IC50 values. GP120 inhibitors: 'n' indicates neutralizing activity, IN inhibitors: 's' indicates integrase strand transfer, '3e' indicates integrase 3'-end processing. '~' indicates approximate values; (6) HIV strain: HIV strains used for evaluation of peptide inhibitory activity. If multiple HIV-1 strains were used, subtype or group information is given in both of the columns (5) and (6). (7) Cell line: cell lines used in the experiments.

Peptide position(1)	Peptide name(2)	Sequence(3)	Target(4)	IC_{50}/EC_{50}/Kd(5)	HIV strain(6)	Cell line(7)	Ref
GP41[638–673]	T20(enfuvirtide)	YTSLIHSLIEESQNQQEKNEQELLELDKWASLWNWF	GP41	B:2.7±0.4nM, IIIB:28nM[109], BCF02>2000nM[109]	B:HXB2, B:IIIB [109], C: BCF02 [109]	293T	[109],[66]
GP41[638–673]	T-20EK	YTSLIEELIKKSEEQQKKNEELKKLEEWAKKWNWF	GP41	B:1.2nM	NL4-3_{D36G}	MT-2	[110]
GP41[621-652]	CP621-652	QIWNNMTWMEWDREINNYTSLIHSLIEESQNQ	GP41	HXB2:8.6±2.5nM, NL4-3:5.8±0.6nM	HXB2,NL4-3	TZM-bl	[111]
GP41 [621-652]	CP32M	VEWNEMTWMEWEREIENYTKLIYKILEESQEQ	GP41	BCF02:10nM[109], IIIB:5nM[109]	IIIB[109], BCF02[109]	MT-2	[109], [112]
GP41[628-661]	NCS-C34-Chol	WMEWK(NCS)REINNYTSLIHSLIEESQNQQEKNEQELLGSGN-Chol	GP41	8.4±2.2nM	HXB2, SF162, CNE28,NL4-3	HEK293T	[113]
GP41[528-581]	17-70	STMGAASMTLTVQARQLLSGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARIL	GP41	391±33nM	HXB2	TZM-bl	[114]
GP41[630-659]	SJ-2176	EWDREINNYTSLIHSLIEESQNQQEKNEQEGGC	GP41	P24-NC:101uM^{E} CPE:142nM^{E} Cell fusion:156nM^{E}	IIIB	MT-2	[115]
GP41[512-544]	IFFA	AVGIGALFLGFLGAAGSTMGARSMTLTVQARQL	GP41		IIIB	SupT1, TF228	[116]
GP41[628-639,641-661]	ABT	WEEWDREINNYT(MPA)LIHELIEESQNQQEKNEQELL	GP41	1.01-66.39nM	NL4-3,subtype A,B,C	TZM-bl	[117]
GP41[626-663]	T1144	TTWEAWDRAIAEYAARIEALLRALQEQQEKNEAALREL	GP41	0.4nM	Bal	TZM-bl	[118]
GP41[638-673][626-663]	TLT35	T20 $(GGGGS)_6$ T1144	GP41	IIIB:11.06±3.12nM, Bal: 2.24±0.68nM,Range: 1.83-27.87nM	IIIB,Bal, Subtype:A,B,C,E,F,G,O	MT-2	[118, 119]
GP41[626-657]	C32-e5.0	Ac-TTWEAWDRAIAEYAARIEALIRAAQEQQEKNC-NH_2	GP41	6.4±1.4nM^{d}			[120, 121]
GP41[626-664]	C39-e5.0	Ac-TTWEAWDRAIAEYAARIEALIRAAQEQQEKNEAELRELC-NH2	GP41	9.9±1.8nM^{d}			[120, 121]
GP41[628-661]	C34	WMEWDREINNYTSLIHSLIEESQNQQEKNEQELL	GP41	>2uM[109]	BCF02[109]	293T	[109],[122, 123]
GP41[636-661]	Aoc−βAla−P26	Aoc−βAla−NNYTSLIHSLIEESQNQQEKNEQELL	GP41	NL4-3D36G:130±12nM, IIIB:14.9±2.99nM	NL4-3D36G, IIIB	MT-2, HL2/3,TZM-b	[124]

GP41[626-661]	MT-C34	MTWMEWDREINNYTSLIHSLIEESQNQQEKNEQELL	GP41	0.5±0.1nM	NL4-3	HL2/3	[125]
GP41[626-649]	MT-SC22EK	MTWEEWDKKIEEYTKKIEELIKKS	GP41	A:3.8-4.6nM B: 1.6-9.3nM C:1.3-10.8nM AE:3.4-15.1nM BC:0.8-6.5nM	A,B,C,A/E,B/C	TZM-bl	[126]
GP41[625-661]	C37	GGHTTWMEWDREINNYTSLIHSLIEESQNQQEKNEQELLGHHHHHH	GP41	HXB2: ~1nM JR-FL:~1.5nM	HXB2,NL4-3,JR-FL, Ba-L	293T	[127]
GP41[628-673]	C46(364H-3L multimer)	WMEWDREINNYTSLIHSLIEESQNQQEKNEQELLELDKWASLWNWF	GP41	JR-FL:23nM, Bal:120nM, 117III: >100nM, HXB2:12nM	JR-FL,Bal, 117III,IIIB,HXB2	U87,293T	[128]
GP41[628-656]	$(Caca29)_2$	$(CAcaWMEWDREINNYTSLIHSLIEESQNQQEKN)_2$	GP41	5.71nM	-	-	[129]
GP41[628-649]	$(CAcaSC22EK)_2$	$(CAcaWEEWDKKIEEYTKKIEELIKKS)_2$	GP41	4.9nM	-	-	[129]
GP41[629-662]	$SC35E(SBn)_5H_9$	Ac-WEEWEKKIHEYTAKIELIKKSEEQQKKNEEELKK-NH_2	GP41	1.02±0.33nM	???	TZM-bl	[130]
GP41[628-673]	V2o	WMTWDREIDNITQTISSAIEESQNQNEKNEQELLKLNQWDIFSNWF	GP41	HXB2:0.42±0.15nM, BaL:0.51±0.11nM,SIVmac251:5.0±3.3nM	HXB2,BaL,SIVmac251	293T	[131]
GP41[553-590]	DP-107	NNLLRAIEAQQHLLQLTVWGIKQLQARILAVERYLKDQ	GP41	2.7uM	IIIB	MT4	[132]
GP41[627-662]	SFT(Sifuvirtide)	SWETWEREIENYTRQIYRILEESQEQQDRNERDLLE	GP41	A:1.81nM, B:10.35nM, C:3.84nM[133]	Subtype A,B,C	MT-2	[133-135]
GP41[559-581]	IQN23	Ac-RMKQIEDKIEEIESKQKKIENEIARIKKL-IEAQQHLLQLTVWGIKQLQARIL-NH_2	GP41	15±7nM	HXB2	293T	[136]
GP41[546-581]	N36	SDIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARIL	GP41	180±70nM[E]	NL4-3	293T, MT-2	[137]
GP41[546-581]	N36Fd	SDIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARIL-GYIPEAPRDGQAYVRKDGEWVLLSTFL	GP41	NL4-3: 56.34±9.24nM,IIIB:99nM; Bal: 182.9nM,93IN101:1.21nM	IIIB,NL4-3, BaL,93IN101	MT-2,PMBC,TZM-bl	[65]
GP41[559-586]	N28Fd	IEAQQHLLQLTVWGIKQLQARILAVERY-GYIPEAPRDGQAYVRKDGEWVLLSTFL	GP41	NL4-3:26.95±0.02nM, IIIB:39nM	IIIB,NL4-3, BaL, 93IN101	MT-2,PMBC,TZM-bl	[65]
GP41[536-581]	N46	TLTVQARQLLSGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARIL	GP41	>1uM (IIIB)	IIIB, 92US657, 94UG103	MT-2,H9,	[138]
GP41[536-581]	N46FdFc	TLTVQARQLLSGIVQQQNNLLRAIEAQQHLLQLTVW-GIKQLQARILGYIPEAPRDGQAYVRKDGEWVLLSTFL-$(H)_6$	GP41	310±25nM	IIIB, 92US657, 94UG103	MT-2,H9,	[138]
GP41[626-663]	T2635	TTWEAWDRAIAEYAARIEALIRAAQEQQEKNEAALREL	GP41	6.24nM~393.0nM	LAI	TZM-bl	[139]
GP41[626-663]	T-2544	MTWEAWDRAIAEYAARIEALIRAAQEQQEKNEAALREL	GP41	7nM	IIIB	PBMC, MT-2	[140]
GP41[626-661]	T-651	MTWMEWDREINNYTSLIHSLIEESQNQQEKNEQELL	GP41	8nM	IIIB	PBMC, MT-2	[140]
GP41[628-683]	P5	WMEWDREINNYTSLIHSLIEESQNQQEKNEQELLELDKWASLWNWFNITNWLWYIK	GP41	~60nM	LAI, JR-CSF	Hela,PBMC	[141]
IN[95-109]	Alpha-1	QETAYFLLKLAGRWP-$CONH_2$	IN	3.5uM		-	[142]
IN[171-187]	Alpha-5	HLKTAVQMAVFIHNFKR-$CONH_2$	IN	3.0uM		-	[142]
IN[196-209]	Alpha-6	AGERIVDIIATDIQ-$CONH_2$	IN	2.0uM		-	[142]
IN[95-107]	Alpha-1s	QETAYFLLKLAGR-$CONH_2$	IN	150uM		-	[142]

IN[196-205]	Alpha-6s	AGERIVDIIA-$CONH_2$	IN	30uM		-	[142]
IN[82-89]	Beta-3	GYIEAEVI-CONH2	IN	>1mM		-	[142]
IN[95-109]	H104	QETAYFLLKLALRWP-CONH2	IN	-	-	-	[143]
IN[97-108]	NL-6	TAYFLLKLAGRW	IN	2.7uM[S], 21 uM[3E]	-	-	[144]
IN[99-104]	NL6-5	YFLLKL	IN	20uM[S, 3E]	-	-	[144]
IN[129-139]	NL-9	ACWWAGIKQEF	IN	56uM[S], 95uM[3E]	-	-	[144]
IN[173-188]	INS K188E	WTAVQMAVFIHNFKRE	IN	5.2 ±0.2uM	HXB2	HEK293T	[145]
IN[92-108]	INH1	ATGQETAYFLLKLAGKA-$CONH_2$	IN	150uM[S], 250uM[3E]	NL4 –3	CEM-12D7	[146]
IN[167-187]	INH5	DQAEHLKTAVQMAVFIHNYKA-$CONH_2$	IN	4.7uM[S], 11uM[3E]			[146]
IN[147-176]	K159	SQGVVESMNKELKKIIGQVRDQAEHLKTAY	IN	16nM[S],16nM[3E]	HXB2D	-	[147],[148],[149]
IN[151-176]	EAA26	VESMNEELKKIIAQVRAQAEHLKTAY	IN	-	-	-	[148],[150]
IN[171-187,196-209]	a5-Cmpi-a6	HLKTAVQMAVFIHNFKR-Cmpi-AGERIVDIIATDIQ-NH2	IN	460 ±30nM			[151]
CA[175-194]	CAC1	Ac-EQASQEVKNWMTETLLVQNA-CONH2	CA	50uM[d]	BH10	-	[100]
CA[207-217]	Capsid1	PAATLEEMMTA	CA			H9	[152]
CA[175-193]	CAC1M	SESAASSVKAWMTETLLVANTSS	CA	8 ±1uM[d]	HXB2	U87-CD4-CXCR4	[153]
CA[175-194]	CAC1C	ESASSSVKAWMTETLLVQNA	CA	19 ±8uM[d]	HXB2	U87-CD4-CXCR4	[153]
CA[178-192]	NYAD-201	AQEVKXWMTXTLLVA (X = (S)-2-alpha- (2'-pentenyl)alanine)	CA	IIIB: 4.29 ±0.62uM, MN: 3.03 ±0.61uM, SF2: 5.06 ±1.37uM, RF: 2.84 ±0.63uM, Bal: 4.73 ±1.92uM, 89.6: 5.21 ±0.87uM	IIIB,MN,RF,SF2,BaL,89.6	MT-2,PBMC	[102]
CA[178-192]	NYAD-202	AQAVKXWMTXTLLVA (X = (S)-alpha- (2'-pentenyl)alanine)	CA	IIIB: 2.36 ±0.33uM, MN: 2.47 ±0.71uM, SF2: 4.48 ±0.84uM, RF: 2.64 ±0.39uM, Bal: 2.23 ±0.44uM, 89.6: 3.471 ±0.22uM	IIIB,MN,RF,SF2,BaL,89.6	MT-2,PBMC	[102]
CA[181-192]	P-1	VKNWMTETLLRQ	CA	3.8 ±3.5uM[d]	BH10	-	[99]
CA[124-133]	peptide 1	IPVGEIYKRW	CA	37 ±10uM[d]			[154]
RT[285-301]	P_{AW}	GTKWLTEWIPLTAEAEC	RT	700 ±200nM[d]	LAI	PBMC	[79]
RT[285-299]	P27	GTKWLTEWIPLTAEC	RT	50 ±10nM[d]	LAI	PBMC	[79]
RT[285-296]	P24	GTKWLTEWIPLC	RT	700 ±50nM[d]	LAI	PBMC	[79]
RT[395-404]	Pep-7	KETWETWWTE	RT	138nM[d]	BH10	-	[77],[155]
RT[389-407]	Peptide1	FKLPIQKETWETWWTEYWE	RT	1.2uM[d]	LAV	MT-4	[78]
PR [83-93]	p-S_8	NIIGRNLLTQI	PR	2.58 ±0.78uM[84]	-	-	[84],[85],[83, 86]
PR[1-5,95-99]	PF1	PQITL-$(G)_3$-CTLNF	PR	40uM(HIV1),20uM(HIV2)	HIV1,HIV2		[82]
MA[71-87]	8L	CH_3CO-GSEELRSLYNTIAVLGC-NH_2	MA	NL4-3: 2.3 ±0.3uM[E] JR-CSF: 7.8uM	NL4-3, JR-CSF	MT-4,PM1/CCR5	[156]
MA[81-97]	9L	CH_3CO-TIAVLYSVHQRIDVKGC-NH_2	MA	NL4-3: 2.1 ±0.5uM[E] JR-CSF: 0.58uM	NL4-3, JR-CSF	MT-4,PM1/CCR5	[156]
MA[47-59]	4/5m	NPGLLETSEGCRQ	MA	615ug/ml	IIIB	H9	[152]
RT [166-185]	4286	KILEPFRKQNPDIVIYQYMD	IN	4.8uM[3E],4.5uM[S]	BH10	-	[157]
RT [516-535]	4321	ELVNQIIEQLIKKEKVYLAW	IN	6.9uM[3E],5uM[S]	BH10	-	[157]
RT [176-195]	34	PDIVIYQYMDDLYVGSDLEI	IN	10uM[S], 6uM[3E]	HXB2	-	
RT [366-385]	53	KQLTEAVQKITTESIVIWGK	IN	7 ±1uM[3E],4 ±1uM[S]	HXB2	-	[158]

RT [396-415]	56	ETWETWWTEYWQATWIPEWE	IN	6 ±1uM^{3E},2 ±1uM^{S}	HXB2	-	[158]
RT [486-505]	65	LQDSGLEVNIVTDSQYALGI	IN	2uM^{S}, 11uM^{3E}	HXB2	-	[158]
RT [526-545]	64	ELVNQIIEQLIKKEKVYLAW	IN	14uM^{S}, 15uM^{3E}	HXB2	-	[158]
IN[46-65]	#4330	KGEAMHGQVDCSPGIWQLDC	RT	4.2 ±0.2uM^{RDDP}, 6.8 ±0.7uM^{DDDP}	HXB2R	-	[159]
Vpr [33-47]	Vpr 33-47	HFPRIWLHSLGQHIY	IN	41uM^{s}, 187uM^{3e}	BH10	-	[90]
Vpr [53-67]	Vpr 53-67	TWAGVEAIIRILQQL	IN	144uM^{s}, >200uM^{3e}	BH10	-	[90]
Vpr [57–71]	Vpr 57-71	VEAIIRILQQLLFIH	RT/IN	0.22uM	BH10	-	[90]
Vpr [61-75]	Vpr 61-75	IRILQQLLFIHFRIG	RT/IN	0.7uM^{rddp}, 1.3uM^{dddp}	BH10	-	[90]
Vpr[55-69]	Vpr-1	AGVEAIIRILQQLLF	IN	-	HXB2, JR-CSF	MT-4	[91]
Vpr[64-75]	Vpr-3 R8	Ac-LQQLLFIHFRIG-RRRRRRRR-NH_2	IN	4 ±0.1nM^{s}[91], 8 ±1nM^{3e}[91], 60 ±10nM^{s}[92], 130 ±20nM^{3e}[92]	HXB2 [91],[92], JR-CSF[91]	MT-4	[91],[92]
Vpr[58-75]	Vpr-4 R8	Ac-EAIIRILQQLLFIHFRIG-RRRRRRRR-NH_2	IN	5 ±2nM^{s}, [91] 6 ±6 nM^{3e}[91] 40 ±10nM^{s}[92] 90 ±10nM^{3e}[92]	HXB2[91],[92], JR-CSF[91]	MT-4[91],[92]	[91],[92]
Vpr [65-79]	Vpr 65-79	QQLLFIHFRIGCQHS	IN	14uM^{s}, 76uM^{3e}	BH10	-	[90]
Vpr[58-75]	Vpr-15	Ac-EAEIRIKQQLLFIHFRIG-RRRRRRRR-NH_2	IN	31 ±10nM^{S}, 40 ±1nM^{3E}	HXB2, JR-CSF	MT-4	[91]
Vif[30-65]	Peptide4	YVSGKARGWFYRHHYESPHPRISSEVHIPLGDARLV	PR	230-250uM	IIIB	Hut 78	[106]
Vif[78-92]	Peptide6	DWHLGQGVSIEWRKK	PR	110uM	IIIB	Hut 78	[106]
Vif[88-98]	Peptide7	EWRKKRYSTQV	PR	25uM[106], 3.31uM[108]	IIIB	Hut 78	[106], [108]
Vif[41-65]	Vif41-65	RHHYESPHPRISSEVHIPLGDARLV	PR/IN		HXB-2	PBL	[104]
p6* [1-8]	TFP	FLREDLAF	PR	98 ±10uM^{I}	HXB2		[160]
p6* [4-6]	-	EDL	PR	50 ±9uM^{I}	HXB2		[160]
PR[1-5]Tat [49-61]p6*[53-56]PR[95-99]	P27	PQITL-RKKRRQRRRPPQV-SFNF- CTLNF	PR	0.23-0.32uM/5uM	A01 patient, LAI[161]	MT-2,H9	[161], [162]
Rev [1-30]	Rev1-30	MAGRSGDSDEELLKTVRLIKFLYQSNPPPS	IN	6.5 ±0.2uM^{d}	HXB2	-	[94]
Rev [13-23]	Rev13-23	LKTVRLIKFLY	IN	2.8 ±0.1uM^{d}	HXB2	HeLa	[94],[96]
Rev [49-74]	Rev49-74	QRQIRSISGWILSTYLGRPAEPVPLQ	IN	11.2 ±0.5uM^{d}	HXB2	-	[94]
Rev [53-67]	Rev53-67	RSISGWILSTYLGRP	IN	6.9 ±0.1uM^{d}	HXB2	HeLa	[94],[96]
CA[229-231]p2[1-3]	6a	RVL-FEA-Nle	PR	NL4-3: 2.60 ±0.4nM, MDR769: 4.40 ±0.7nM	NL4-3, MDR769	-	[163]
GP120[280-302]	NTM	RSANFTDNAKTIIVQLNQSVEIN	CD4 receptor		BH-10	-	[164]
GP120[424-433]	Peptide 1	INMWQEVGKA	CD4	28uM	IIIB	-	[165]
GP120[365-373]	Peptide 2	SGGDPEIVT	CD4	6uM	IIIB	-	[165]
GP120[312-317]	SPC3	[GPGRAF]8-K_4-K_2-K-βA	chemokine receptors α/β	7.7 ±0.4uM	LAI	Xenopus oocyte	[166]
GP120[293-334]	V3-BH10	EINCTRPNNNTRKSIRIQRGPGRAFVTIGKIGNMRQAHCNIS	IgG, MAbs,CD19	-	IIIB,MN	MT-4, PMBC	[167]
GP120[290-334]	V3-89.6	ESVVINCTRPNNNTRRRLSIGPGRAFYARRNIIGDIRQAHCNIS	IgG, MAbs,CD19	-	IIIB,MN	MT-4, PMBC	[167]
GP120[290-320,323-334]	V3-ELI	ESVKITCARPYQNTRQRTPIGLGQSLYTTRSRSIIGQAHCNIS	IgG, MAbs,CD19	-	IIIB,MN	MT-4, PMBC	[167]

GP120[298-321]	62.19	RPNNNTRKRIRIQRGPGRAFVAIE	F39F,447-52D Fab	31%[n]	IIIB,89.6	MT-2	[168]
GP120[296-313,315-331]	$V3_B$-FP	CTRPNNNTRKSIRIGPGQTFYATGDIIGDIRQAHC		>50%[n]	BZ167, DJ263, NL43		[169]
GP120[421-436,298-321]	C4–V3 T303C–E322C	KQIINMWQEVGKAMYA-RPNNNCRKSIHIGPGRAFYTTGCG	chemokine receptors	-	IIIB,NL4-3, JRFL	293T,Rabbit	[63]
GP120[326-340]	15K	IRKAHCNISRAKWND	CXCR4,CCR5			293T,PBMC, MDM	[170]
GP120[326-340]	15D	IRKAHCNISRADWND	CXCR4,CCR5			293T,PBMC, MDM	[170]
GP120[157-171]	CT319	CSFNITTEIRDKVKK	Tat		HXB2	U937	[171]
IN[170-191], IN[214-228], IN[259-273]	$CCD_{170-191}$,$CTD_{214-228}$,$CTD_{259-274}$	(1)EHLKTAVQMAVFIHNFKRKGGI,(2)QKQITKIQNFRVYYR (3)VVPRRKVKIIRDYGK	Transportin-SR2	-	-	-	[172]
IN[161-174]	NLS(IN)	IIGQVRDQAEHLKC-NH2	Importin-alpha		-		[173]
Tat[48-57]	R10	RRRRRRRRRR		50uM[e]	IIIB	MAGI	[87]
Tat[49-58]	Tat11	RKKRRGRRRC-NH2		5nM[d]		Colo-205	[88]
Tat[11-50]	-	WKHPGSQPKTACTNCYCKKCCFHCQVCFITKALGISYGRK	CXCR4		NL4-3	293T	[174, 175]
Rev[34-47]	Rev 8	Ac-RRRRERQRKRRRRR-OH	RRE	~150nM	-	-	[93]
MA[11-47]	p17(11-47)	GELDRWEKIRLRPGGKKKYKLKHIVWASRELERFAVN	Ca2+/CaM	-	-	-	[137]
GP41[577-586]		QARVLAVERY	IgA		IIIB,ADA	TZM-bl	[176]
GP41[628–683]	Peptide P5	WMEWDREINNYTSLIHSLIEESQNQQEKNEQELLELDKWASLWNWFNITNWLWYIK	GP41-Antibody	61±1.5uM	HXB2	Hela-CD4-LTR-LacZ	[141]

3.8 Additional file 3: Software

This document briefly describes the manual of functions and algorithms included in our software. We present the software for full-length genome alignment, the intra- and inter-clade genome diversity analysis and the data visualization. Our example datasets are available in the toolbox package.

Motivation: Extensive full-length sequences of HIV genomes have been accumulated in the past few years. To our best knowledge, a toolbox developed for analyzing full-length HIV genome has not been reported. Here, we offer a Matlab toolbox for HIV full-length genomic analysis. We provide an alignment tool that optimizes HIV amino acid alignments given the input files of nucleotides genomic sequences. Classical alignment tools either perform nucleotide or amino acid alignments. In order words, nucleotide (amino acid) sequences are optimized when nucleotide (amino acid)

sequences are inputs. Most classical tools do not optimize amino acid alignments based on nucleotide genomic sequences, because of the overlapping regions in three open reading frames. Considering different lengths and locations of 15 HIV proteins in 3 open reading frames of the HIV genome, our tool has been developed to optimize the genomic alignment using the reference mapping strategy.

Protocol: HIV-1 and HIV-2 genome encodes 15 proteins in three open reading frames. The number of nucleotides in one HIV genomic sequence is usually between 8500 and 8800. We have developed a tool to improve codon sequence alignments. Briefly, the input of our toolbox requires the nucleotide sequence alignment, which can be prepared using many classical nucleotide alignment tools (e.g. SeaView [31], Mega [177], Mafft [30]). Next, our toolbox uses the reference genome to assemble the nucleotide sequence alignment from each protein coding region. Thereafter, the alignments in protein regions are optimized by maximizing the matched codons based on the amino acid substitution matrix.

We first describe our algorithm for improving the amino acid alignments from nucleotide sequences in one protein coding region.

Algorithm: Alignment of codon sequences
Input: One MSA file with nucleotide genomic sequences
Output: One MSA file with aligned amino acid sequences
Step 1: Arrange the nucleotide positions into codon triples. /*--A-B---C--- => --ABC------- */
Step 2: Refine small nucleotides into codon columns. /*-AAA--BBB- => --AAA--BBB-- */
/*-AAA-BBB-- => --AAA--BBB-- */
Step 3: Refine codon positions to the left or right side if substitution scores are improved.
Step 4: Optimize the codon alignments at the insertion and deletion regions.
Step 5: Transform codon sequence alignments into amino acid alignments.

In our toolbox, the function called "TransferNucleotide2AminoAcidAlignment.m" implements the above algorithm. A simple example is given to run the alignment tool. Note that MSAInputFile is the path of the MSA input file and MSAOutputFile is the path of the MSA output file.

Example:

```
[ Seq,SeqTitle,SeqID ] = ExtractSequenceOut(  MSAInputFile  ); % collect the nucleotide input
Seq = TransferNucleotide2AminoAcidAlignment( Seq,SeqTitle );%optimize codon alignments
[a,b] = WriteSequence2Fasta( Seq, SeqTitle, MSAOutputFile  ); % output improved alignments
```

Secondly, six steps are performed to align full-length HIV genomic sequences. An example is provided in the file called "StartFile_FullGnomeAlignmentTool.m".

Procedure for HIV genomic alignment:
(1) We collect the information of HXB2 genome reference.
(2) We concatenate each HIV-1 protein region in the full-length genome based on HXB2.
(3) We perform the codon sequence alignment of HIV-1 proteins using full-length alignment tool.
(4) We examine sequence quality using other sequence visualization software, such as Seaview.
(5) We transform the sequences from nucleotide forms to amino acid forms.
(6) We assemble the concatenated protein sequences into full-length amino acid genomes.

Thirdly, "AnalysisGenomeIntraSubtypeDiversity.m" implements the computation of the intra-clade genomic diversity. "AnalysisGenomeInterSubtypeDiversity.m" implements the computation of the inter-clade genomic diversity.

We implemented our toolbox in Matlab 2013a under the Linux system. A simple example to use our toolbox is provided in: "MeasureGenomeGeneticDiversity.m". If you have encountered the difficulty of using the toolbox in different Matlab versions or computer systems, please consult Guangdi Li (liguangdi.research@gmail.com).

3.9 References

1. Hemelaar J. The origin and diversity of the HIV-1 pandemic. *Trends Mol Med* 2012,**18**:182-192.
2. Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, *et al.* Diversity considerations in HIV-1 vaccine selection. *Science* 2002,**296**:2354-2360.
3. Frankel AD, Young JA. HIV-1: fifteen proteins and an RNA. *Annu Rev Biochem* 1998,**67**:1-25.
4. Engelman A, Cherepanov P. The structural biology of HIV-1: mechanistic and therapeutic insights. *Nat Rev Microbiol* 2012,**10**:279-290.
5. Acosta EG, Kumar A, Bartenschlager R. Revisiting dengue virus-host cell interaction: new insights into molecular and cellular virology. *Adv Virus Res* 2014,**88**:1-109.
6. Perry CM. Elvitegravir/cobicistat/emtricitabine/tenofovir disoproxil fumarate single-tablet regimen (Stribild(R)): a review of its use in the management of HIV-1 infection in adults. *Drugs* 2014,**74**:75-97.
7. Tilton JC, Doms RW. Entry inhibitors in the treatment of HIV-1 infection. *Antiviral Res* 2010,**85**:91-100.
8. Fauci AS, Folkers GK, Dieffenbach CW. HIV-AIDS: much accomplished, much to do. *Nat Immunol* 2013,**14**:1104-1107.
9. Li G, Verheyen J, Rhee SY, Voet A, Vandamme AM, Theys K. Functional conservation of HIV-1 gag: implications for rational drug design. *Retrovirology* 2013,**10**:126.
10. Stephenson KE, Barouch DH. A global approach to HIV-1 vaccine development. *Immunol Rev* 2013,**254**:295-304.
11. Rolland M, Tovanabutra S, deCamp AC, Frahm N, Gilbert PB, Sanders-Buell E, *et al.* Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. *Nat Med* 2011,**17**:366-371.
12. Rolland M, Edlefsen PT, Larsen BB, Tovanabutra S, Sanders-Buell E, Hertz T, *et al.* Increased HIV-1 vaccine efficacy against viruses with genetic signatures in Env V2. *Nature* 2012,**490**:417-420.
13. Arien KK, Vanham G, Arts EJ. Is HIV-1 evolving to a less virulent form in humans? *Nat Rev Microbiol* 2007,**5**:141-151.

14. Herbeck JT, Rolland M, Liu Y, McLaughlin S, McNevin J, Zhao H, *et al.* Demographic processes affect HIV-1 evolution in primary infection before the onset of selective processes. *J Virol* 2011,**85**:7523-7534.
15. Rousseau CM, Birditt BA, McKay AR, Stoddard JN, Lee TC, McLaughlin S, *et al.* Large-scale amplification, cloning and sequencing of near full-length HIV-1 subtype C genomes. *J Virol Methods* 2006,**136**:118-125.
16. Wang YE, Li B, Carlson JM, Streeck H, Gladden AD, Goodman R, *et al.* Protective HLA class I alleles that restrict acute-phase CD8+ T-cell responses are associated with viral escape mutations located in highly conserved regions of human immunodeficiency virus type 1. *J Virol* 2009,**83**:1845-1855.
17. Brown BK, Darden JM, Tovanabutra S, Oblander T, Frost J, Sanders-Buell E, *et al.* Biologic and genetic characterization of a panel of 60 human immunodeficiency virus type 1 isolates, representing clades A, B, C, D, CRF01_AE, and CRF02_AG, for the development and assessment of candidate vaccines. *J Virol* 2005,**79**:6089-6101.
18. Fernandez-Garcia A, Cuevas MT, Munoz-Nieto M, Ocampo A, Pinilla M, Garcia V, *et al.* Development of a panel of well-characterized human immunodeficiency virus type 1 isolates from newly diagnosed patients including acute and recent infections. *AIDS Res Hum Retroviruses* 2009,**25**:93-102.
19. Kousiappa I, Van De Vijver DA, Kostrikis LG. Near full-length genetic analysis of HIV sequences derived from Cyprus: evidence of a highly polyphyletic and evolving infection. *AIDS Res Hum Retroviruses* 2009,**25**:727-740.
20. Sanabani SS, Pessoa R, Soares de Oliveira AC, Martinez VP, Giret MT, de Menezes Succi RC, *et al.* Variability of HIV-1 genomes among children and adolescents from Sao Paulo, Brazil. *PLoS One* 2013,**8**:e62552.
21. Fernandez-Garcia A, Revilla A, Vazquez-de Parga E, Vinogradova A, Rakhmanova A, Karamov E, *et al.* The analysis of near full-length genome sequences of HIV type 1 subtype A viruses from Russia supports the monophyly of major intrasubtype clusters. *AIDS Res Hum Retroviruses* 2012,**28**:1340-1343.
22. Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, *et al.* Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog* 2012,**8**:e1002529.
23. Guyader M, Emerman M, Sonigo P, Clavel F, Montagnier L, Alizon M. Genome organization and transactivation of the human immunodeficiency virus type 2. *Nature* 1987,**326**:662-669.
24. Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C, Detours V. Evolutionary and immunological implications of contemporary HIV-1 variation. *Br Med Bull* 2001,**58**:19-42.
25. van der Kuyl AC, Berkhout B. The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus. *Retrovirology* 2012,**9**:92.
26. Snoeck J, Fellay J, Bartha I, Douek DC, Telenti A. Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints. *Retrovirology* 2011,**8**:87.
27. Mayrose I, Stern A, Burdelova EO, Sabo Y, Laham-Karam N, Zamostiano R, *et al.* Synonymous site conservation in the HIV-1 genome. *BMC Evol Biol* 2013,**13**:164.
28. Santoro MM, Perno CF. HIV-1 Genetic Variability and Clinical Implications. *ISRN Microbiol* 2013,**2013**:481314.
29. Hemelaar J, Gouws E, Ghys PD, Osmanov S, Isolation W-UNfH, Characterisation. Global trends in molecular epidemiology of HIV-1 during 2000-2007. *AIDS* 2011,**25**:679-689.
30. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004,**32**:1792-1797.
31. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 2010,**27**:221-224.
32. Hooft RW, Vriend G, Sander C, Abola EE. Errors in protein structures. *Nature* 1996,**381**:272.
33. Pinney JW, Dickerson JE, Fu W, Sanders-Beer BE, Ptak RG, Robertson DL. HIV-host interactions: a map of viral perturbation of the host system. *AIDS* 2009,**23**:549-554.
34. Llano A, Frahm N, Brander C. How to optimally define optimal cytotoxic T lymphocyte epitopes in HIV infection. *HIV molecular immunology* 2009,**2009**:3-24.
35. Pornillos O, Ganser-Pornillos BK, Kelly BN, Hua Y, Whitby FG, Stout CD, *et al.* X-ray structures of the hexameric building block of the HIV capsid. *Cell* 2009,**137**:1282-1292.
36. Daugherty MD, Liu B, Frankel AD. Structural basis for cooperative RNA binding and export complex assembly by HIV Rev. *Nat Struct Mol Biol* 2010,**17**:1337-1342.

37. Auclair JR, Green KM, Shandilya S, Evans JE, Somasundaran M, Schiffer CA. Mass spectrometry analysis of HIV-1 Vif reveals an increase in ordered structure upon oligomerization in regions necessary for viral infectivity. *Proteins* 2007,**69**:270-284.
38. Buchan DW, Ward SM, Lobley AE, Nugent TC, Bryson K, Jones DT. Protein annotation and modelling servers at University College London. *Nucleic Acids Res* 2010,**38**:W563-568.
39. Klose DP, Wallace BA, Janes RW. 2Struc: the secondary structure server. *Bioinformatics* 2010,**26**:2624-2625.
40. Xue B, Mizianty MJ, Kurgan L, Uversky VN. Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. *Cell Mol Life Sci* 2012,**69**:1211-1259.
41. Ishida T, Kinoshita K. Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics* 2008,**24**:1344-1348.
42. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* 2006,**7**:208.
43. Deng X, Eickholt J, Cheng J. PreDisorder: ab initio sequence-based prediction of protein disordered regions. *BMC Bioinformatics* 2009,**10**:436.
44. Levy ED. A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J Mol Biol* 2010,**403**:660-670.
45. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* 2004,**25**:1605-1612.
46. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 2010,**5**:e9490.
47. Spira S, Wainberg MA, Loemba H, Turner D, Brenner BG. Impact of clade diversity on HIV-1 virulence, antiretroviral drug sensitivity and drug resistance. *J Antimicrob Chemother* 2003,**51**:229-240.
48. Haynes BF, Gilbert PB, McElrath MJ, Zolla-Pazner S, Tomaras GD, Alam SM, *et al.* Immune-correlates analysis of an HIV-1 vaccine efficacy trial. *N Engl J Med* 2012,**366**:1275-1286.
49. Rambaut A, Posada D, Crandall KA, Holmes EC. The causes and consequences of HIV evolution. *Nat Rev Genet* 2004,**5**:52-61.
50. Falkowski PG, Fenchel T, Delong EF. The microbial engines that drive Earth's biogeochemical cycles. *Science* 2008,**320**:1034-1039.
51. Dunwell JM, Culham A, Carter CE, Sosa-Aguirre CR, Goodenough PW. Evolution of functional diversity in the cupin superfamily. *Trends Biochem Sci* 2001,**26**:740-746.
52. Buck PM, Kumar S, Singh SK. On the role of aggregation prone regions in protein evolution, stability, and enzymatic catalysis: insights from diverse analyses. *PLoS Comput Biol* 2013,**9**:e1003291.
53. Zolla-Pazner S, Cardozo T. Structure-function relationships of HIV-1 envelope sequence-variable regions refocus vaccine design. *Nat Rev Immunol* 2010,**10**:527-535.
54. Jager S, Cimermancic P, Gulbahce N, Johnson JR, McGovern KE, Clarke SC, *et al.* Global landscape of HIV-human protein complexes. *Nature* 2012,**481**:365-370.
55. Burnett JC, Zaia JA, Rossi JJ. Creating genetic resistance to HIV. *Curr Opin Immunol* 2012,**24**:625-632.
56. He Y. Synthesized peptide inhibitors of HIV-1 gp41-dependent membrane fusion. *Curr Pharm Des* 2013,**19**:1800-1809.
57. Roberts KE, Cushing PR, Boisguerin P, Madden DR, Donald BR. Computational design of a PDZ domain peptide inhibitor that rescues CFTR activity. *PLoS Comput Biol* 2012,**8**:e1002477.
58. Durrant JD, McCammon JA. Molecular dynamics simulations and drug discovery. *BMC Biol* 2011,**9**:71.
59. Liu S, Lu H, Niu J, Xu Y, Wu S, Jiang S. Different from the HIV fusion inhibitor C34, the anti-HIV drug Fuzeon (T-20) inhibits HIV-1 entry by targeting multiple sites in gp41 and gp120. *J Biol Chem* 2005,**280**:11259-11273.
60. Rolland M, Nickle DC, Mullins JI. HIV-1 group M conserved elements vaccine. *PLoS Pathog* 2007,**3**:e157.
61. Kwong PD, Mascola JR, Nabel GJ. Broadly neutralizing antibodies and the search for an HIV-1 vaccine: the end of the beginning. *Nat Rev Immunol* 2013,**13**:693-701.
62. Ghaffari G, Tuttle DL, Briggs D, Burkhardt BR, Bhatt D, Andiman WA, *et al.* Complex determinants in human immunodeficiency virus type 1 envelope gp120 mediate CXCR4-dependent infection of macrophages. *J Virol* 2005,**79**:13250-13261.

63. Moseri A, Tantry S, Sagi Y, Arshava B, Naider F, Anglister J. An optimally constrained V3 peptide is a better immunogen than its linear homolog or HIV-1 gp120. *Virology* 2010,**401**:293-304.
64. Frey G, Chen J, Rits-Volloch S, Freeman MM, Zolla-Pazner S, Chen B. Distinct conformational states of HIV-1 gp41 are recognized by neutralizing and non-neutralizing antibodies. *Nat Struct Mol Biol* 2010,**17**:1486-1491.
65. Chen X, Lu L, Qi Z, Lu H, Wang J, Yu X, *et al.* Novel recombinant engineered gp41 N-terminal heptad repeat trimers and their potential as anti-HIV-1 therapeutics or microbicides. *J Biol Chem* 2010,**285**:25506-25515.
66. Champagne K, Shishido A, Root MJ. Interactions of HIV-1 inhibitory peptide T20 with the gp41 N-HR coiled coil. *J Biol Chem* 2009,**284**:3619-3627.
67. Welch BD, VanDemark AP, Heroux A, Hill CP, Kay MS. Potent D-peptide inhibitors of HIV-1 entry. *Proc Natl Acad Sci U S A* 2007,**104**:16828-16833.
68. Pang W, Tam SC, Zheng YT. Current peptide HIV type-1 fusion inhibitors. *Antivir Chem Chemother* 2009,**20**:1-18.
69. Ashkenazi A, Shai Y. Insights into the mechanism of HIV-1 envelope induced membrane fusion as revealed by its inhibitory peptides. *Eur Biophys J* 2011,**40**:349-357.
70. Welch BD, Francis JN, Redman JS, Paul S, Weinstock MT, Reeves JD, *et al.* Design of a potent D-peptide HIV-1 entry inhibitor with a strong barrier to resistance. *J Virol* 2010,**84**:11235-11244.
71. Eckert DM, Malashkevich VN, Hong LH, Carr PA, Kim PS. Inhibiting HIV-1 entry: discovery of D-peptide inhibitors that target the gp41 coiled-coil pocket. *Cell* 1999,**99**:103-115.
72. Poeschla EM. Integrase, LEDGF/p75 and HIV replication. *Cell Mol Life Sci* 2008,**65**:1403-1424.
73. Craigie R, Bushman FD. HIV DNA Integration. *Cold Spring Harb Perspect Med* 2012,**2**:a006890.
74. Maes M, Loyter A, Friedler A. Peptides that inhibit HIV-1 integrase by blocking its protein-protein interactions. *FEBS J* 2012,**279**:2795-2809.
75. La Regina G, Coluccia A, Silvestri R. Looking for an active conformation of the future HIV type-1 non-nucleoside reverse transcriptase inhibitors. *Antivir Chem Chemother* 2010,**20**:213-237.
76. Huang H, Chopra R, Verdine GL, Harrison SC. Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance. *Science* 1998,**282**:1669-1675.
77. Depollier J, Hourdou ML, Aldrian-Herrada G, Rothwell P, Restle T, Divita G. Insight into the mechanism of a peptide inhibitor of HIV reverse transcriptase dimerization. *Biochemistry* 2005,**44**:1909-1918.
78. Divita G, Restle T, Goody RS, Chermann JC, Baillon JG. Inhibition of human immunodeficiency virus type 1 reverse transcriptase dimerization using synthetic peptides derived from the connection domain. *J Biol Chem* 1994,**269**:13080-13083.
79. Agopian A, Gros E, Aldrian-Herrada G, Bosquet N, Clayette P, Divita G. A new generation of peptide-based inhibitors targeting HIV-1 reverse transcriptase conformational flexibility. *J Biol Chem* 2009,**284**:254-264.
80. Ishima R, Torchia DA, Louis JM. Mutational and structural studies aimed at characterizing the monomer of HIV-1 protease and its precursor. *J Biol Chem* 2007,**282**:17190-17199.
81. Sperka T, Boross P, Eizert H, Tozser J, Bagossi P. Effect of mutations on the dimer stability and the pH optimum of the human foamy virus protease. *Protein Eng Des Sel* 2006,**19**:369-375.
82. Bowman MJ, Chmielewski J. Novel strategies for targeting the dimerization interface of HIV protease with cross-linked interfacial peptides. *Biopolymers* 2002,**66**:126-133.
83. Bonomi M, Barducci A, Gervasio FL, Parrinello M. Multiple routes and milestones in the folding of HIV-1 protease monomer. *PLoS One* 2010,**5**:e13208.
84. Broglia RA, Provasi D, Vasile F, Ottolina G, Longhi R, Tiana G. A folding inhibitor of the HIV-1 protease. *Proteins* 2006,**62**:928-933.
85. Bonomi M, Gervasio FL, Tiana G, Provasi D, Broglia RA, Parrinello M. Insight into the folding inhibition of the HIV-1 protease by a small peptide. *Biophys J* 2007,**93**:2813-2821.
86. Broglia RA, Tiana G, Sutto L, Provasi D, Simona F. Design of HIV-1-PR inhibitors that do not create resistance: blocking the folding of single monomers. *Protein Sci* 2005,**14**:2668-2681.

87. Keogan S, Passic S, Krebs FC. Infection by CXCR4-Tropic Human Immunodeficiency Virus Type 1 Is Inhibited by the Cationic Cell-Penetrating Peptide Derived from HIV-1 Tat. *Int J Pept* 2012,**2012**:349427.
88. Friedler A, Friedler D, Luedtke NW, Tor Y, Loyter A, Gilon C. Development of a functional backbone cyclic mimetic of the HIV-1 Tat arginine-rich motif. *J Biol Chem* 2000,**275**:23783-23789.
89. Davidson A, Leeper TC, Athanassiou Z, Patora-Komisarska K, Karn J, Robinson JA, *et al.* Simultaneous recognition of HIV-1 TAR RNA bulge and loop sequences by cyclic peptide mimics of Tat protein. *Proc Natl Acad Sci U S A* 2009,**106**:11931-11936.
90. Gleenberg IO, Herschhorn A, Hizi A. Inhibition of the activities of reverse transcriptase and integrase of human immunodeficiency virus type-1 by peptides derived from the homologous viral protein R (Vpr). *J Mol Biol* 2007,**369**:1230-1243.
91. Suzuki S, Maddali K, Hashimoto C, Urano E, Ohashi N, Tanaka T, *et al.* Peptidic HIV integrase inhibitors derived from HIV gene products: structure-activity relationship studies. *Bioorg Med Chem* 2010,**18**:6771-6775.
92. Suzuki S, Urano E, Hashimoto C, Tsutsumi H, Nakahara T, Tanaka T, *et al.* Peptide HIV-1 integrase inhibitors from HIV-1 gene products. *J Med Chem* 2010,**53**:5356-5360.
93. Mills NL, Daugherty MD, Frankel AD, Guy RK. An alpha-helical peptidomimetic inhibitor of the HIV-1 Rev-RRE interaction. *J Am Chem Soc* 2006,**128**:3496-3497.
94. Rosenbluh J, Hayouka Z, Loya S, Levin A, Armon-Omer A, Britan E, *et al.* Interaction between HIV-1 Rev and integrase proteins: a basis for the development of anti-HIV peptides. *J Biol Chem* 2007,**282**:15743-15753.
95. Levin A, Hayouka Z, Friedler A, Loyter A. Nucleocytoplasmic shuttling of HIV-1 integrase is controlled by the viral Rev protein. *Nucleus* 2010,**1**:190-201.
96. Levin A, Rosenbluh J, Hayouka Z, Friedler A, Loyter A. Integration of HIV-1 DNA is regulated by interplay between viral rev and cellular LEDGF/p75 proteins. *Mol Med* 2010,**16**:34-44.
97. Hayouka Z, Rosenbluh J, Levin A, Maes M, Loyter A, Friedler A. Peptides derived from HIV-1 Rev inhibit HIV-1 integrase in a shiftide mechanism. *Biopolymers* 2008,**90**:481-487.
98. Levin A, Hayouka Z, Helfer M, Brack-Werner R, Friedler A, Loyter A. Peptides derived from HIV-1 integrase that bind Rev stimulate viral genome integration. *PLoS One* 2009,**4**:e4155.
99. Domenech R, Bocanegra R, Gonzalez-Muniz R, Gomez J, Mateu MG, Neira JL. Larger helical populations in peptides derived from the dimerization helix of the capsid protein of HIV-1 results in peptide binding toward regions other than the "hotspot" interface. *Biomacromolecules* 2011,**12**:3252-3264.
100. Garzon MT, Lidon-Moya MC, Barrera FN, Prieto A, Gomez J, Mateu MG, *et al.* The dimerization domain of the HIV-1 capsid protein binds a capsid protein-derived peptide: a biophysical characterization. *Protein Sci* 2004,**13**:1512-1523.
101. Zhang H, Zhao Q, Bhattacharya S, Waheed AA, Tong X, Hong A, *et al.* A cell-penetrating helical peptide as a potential HIV-1 inhibitor. *J Mol Biol* 2008,**378**:565-580.
102. Zhang H, Curreli F, Zhang X, Bhattacharya S, Waheed AA, Cooper A, *et al.* Antiviral activity of alpha-helical stapled peptides designed from the HIV-1 capsid dimerization domain. *Retrovirology* 2011,**8**:28.
103. Sticht J, Humbert M, Findlow S, Bodem J, Muller B, Dietrich U, *et al.* A peptide inhibitor of HIV-1 assembly in vitro. *Nat Struct Mol Biol* 2005,**12**:671-677.
104. Potash MJ, Bentsman G, Muir T, Krachmarov C, Sova P, Volsky DJ. Peptide inhibitors of HIV-1 protease and viral infection of peripheral blood lymphocytes based on HIV-1 Vif. *Proc Natl Acad Sci U S A* 1998,**95**:13865-13868.
105. Adekale MA, Cane PA, McCrae MA. Changes in the Vif protein of HIV-1 associated with the development of resistance to inhibitors of viral protease. *J Med Virol* 2005,**75**:195-201.
106. Baraz L, Friedler A, Blumenzweig I, Nussinuv O, Chen N, Steinitz M, *et al.* Human immunodeficiency virus type 1 Vif-derived peptides inhibit the viral protease and arrest virus production. *FEBS Lett* 1998,**441**:419-426.
107. Baraz L, Hutoran M, Blumenzweig I, Katzenellenbogen M, Friedler A, Gilon C, *et al.* Human immunodeficiency virus type 1 Vif binds the viral protease by interaction with its N-terminal region. *J Gen Virol* 2002,**83**:2225-2230.
108. Friedler A, Blumenzweig I, Baraz L, Steinitz M, Kotler M, Gilon C. Peptides derived from HIV-1 Vif: a non-substrate based novel type of HIV-1 protease inhibitors. *J Mol Biol* 1999,**287**:93-101.

109. He Y, Cheng J, Lu H, Li J, Hu J, Qi Z, *et al.* Potent HIV fusion inhibitors against Enfuvirtide-resistant HIV-1 strains. *Proc Natl Acad Sci U S A* 2008,**105**:16332-16337.
110. Shimane K, Kawaji K, Miyamoto F, Oishi S, Watanabe K, Sakagami Y, *et al.* HIV-1 resistance mechanism to an electrostatically constrained peptide fusion inhibitor that is active against T-20-resistant strains. *Antimicrob Agents Chemother* 2013.
111. Chong H, Yao X, Qiu Z, Qin B, Han R, Waltersperger S, *et al.* Discovery of critical residues for viral entry and inhibition through structural Insight of HIV-1 fusion inhibitor CP621-652. *J Biol Chem* 2012,**287**:20281-20289.
112. Yao X, Chong H, Zhang C, Qiu Z, Qin B, Han R, *et al.* Structural Basis of Potent and Broad HIV-1 Fusion Inhibitor CP32M. *J Biol Chem* 2012,**287**:26618-26629.
113. Zhao L, Tong P, Chen YX, Hu ZW, Wang K, Zhang YN, *et al.* A multi-functional peptide as an HIV-1 entry inhibitor based on self-concentration, recognition, and covalent attachment. *Org Biomol Chem* 2012,**10**:6512-6520.
114. Sackett K, Wexler-Cohen Y, Shai Y. Characterization of the HIV N-terminal fusion peptide-containing region in context of key gp41 fusion conformations. *J Biol Chem* 2006,**281**:21755-21762.
115. Jiang S, Lin K, Strick N, Neurath AR. HIV-1 inhibition by a peptide. *Nature* 1993,**365**:113.
116. Gerber D, Pritsker M, Gunther-Ausborn S, Johnson B, Blumenthal R, Shai Y. Inhibition of HIV-1 envelope glycoprotein-mediated cell fusion by a DL-amino acid-containing fusion peptide: possible recognition of the fusion complex. *J Biol Chem* 2004,**279**:48224-48230.
117. Chong H, Yao X, Zhang C, Cai L, Cui S, Wang Y, *et al.* Biophysical property and broad anti-HIV activity of albuvirtide, a 3-maleimimidopropionic acid-modified peptide fusion inhibitor. *PLoS One* 2012,**7**:e32599.
118. Cai L, Pan C, Xu L, Shui Y, Liu K, Jiang S. Interactions between different generation HIV-1 fusion inhibitors and the putative mechanism underlying the synergistic anti-HIV-1 effect resulting from their combination. *FASEB J* 2012,**26**:1018-1026.
119. Pan C, Cai L, Lu H, Lu L, Jiang S. A novel chimeric protein-based HIV-1 fusion inhibitor targeting gp41 glycoprotein with high potency and stability. *J Biol Chem* 2011,**286**:28425-28434.
120. Cai L, Balogh E, Gochin M. Stable extended human immunodeficiency virus type 1 gp41 coiled coil as an effective target in an assay for high-affinity fusion inhibitors. *Antimicrob Agents Chemother* 2009,**53**:2444-2449.
121. Gochin M. A Suite of Modular Fluorescence Assays Interrogate the Human Immunodeficiency Virus Glycoprotein-41 Coiled Coil and Assist in Determining Binding Mechanism of Low Molecular Weight Fusion Inhibitors. *Assay Drug Dev Technol* 2012.
122. Lu M, Blacklow SC, Kim PS. A trimeric structural domain of the HIV-1 transmembrane glycoprotein. *Nat Struct Biol* 1995,**2**:1075-1082.
123. Hollmann A, Matos PM, Augusto MT, Castanho MA, Santos NC. Conjugation of cholesterol to HIV-1 fusion inhibitor C34 increases peptide-membrane interactions potentiating its action. *PLoS One* 2013,**8**:e60302.
124. Wang C, Shi W, Cai L, Lu L, Wang Q, Zhang T, *et al.* Design, synthesis, and biological evaluation of highly potent small molecule-peptide conjugates as new HIV-1 fusion inhibitors. *J Med Chem* 2013,**56**:2527-2539.
125. Chong H, Yao X, Sun J, Qiu Z, Zhang M, Waltersperger S, *et al.* The M-T hook structure is critical for design of HIV-1 fusion inhibitors. *J Biol Chem* 2012,**287**:34558-34568.
126. Chong H, Yao X, Qiu Z, Sun J, Zhang M, Waltersperger S, *et al.* Short-peptide fusion inhibitors with high potency against wild-type and enfuvirtide-resistant HIV-1. *FASEB J* 2013,**27**:1203-1213.
127. Kahle KM, Steger HK, Root MJ. Asymmetric deactivation of HIV-1 gp41 following fusion inhibitor binding. *PLoS Pathog* 2009,**5**:e1000674.
128. Dervillez X, Huther A, Schuhmacher J, Griesinger C, Cohen JH, von Laer D, *et al.* Stable expression of soluble therapeutic peptides in eukaryotic cells by multimerisation: application to the HIV-1 fusion inhibitory peptide C46. *ChemMedChem* 2006,**1**:330-339.
129. Ling Y, Xue H, Jiang X, Cai L, Liu K. Increase of anti-HIV activity of C-peptide fusion inhibitors using a bivalent drug design approach. *Bioorg Med Chem Lett* 2013,**23**:4770-4773.
130. Bai Y, Xue H, Wang K, Cai L, Qiu J, Bi S, *et al.* Covalent fusion inhibitors targeting HIV-1 gp41 deep pocket. *Amino Acids* 2013,**44**:701-713.
131. Brauer F, Schmidt K, Zahn RC, Richter C, Radeke HH, Schmitz JE, *et al.* A rationally engineered anti-HIV peptide fusion inhibitor with greatly reduced immunogenicity. *Antimicrob Agents Chemother* 2013,**57**:679-688.

132. Kazmierski WM, Hazen RJ, Aulabaugh A, StClair MH. Inhibitors of human immunodeficiency virus type 1 derived from gp41 transmembrane protein: structure--activity studies. *J Med Chem* 1996,**39**:2681-2689.
133. Yao X, Chong H, Zhang C, Waltersperger S, Wang M, Cui S, *et al.* Broad antiviral activity and crystal structure of HIV-1 fusion inhibitor sifuvirtide. *J Biol Chem* 2012,**287**:6788-6796.
134. He Y, Xiao Y, Song H, Liang Q, Ju D, Chen X, *et al.* Design and evaluation of sifuvirtide, a novel HIV-1 fusion inhibitor. *J Biol Chem* 2008,**283**:11126-11134.
135. Wang RR, Yang LM, Wang YH, Pang W, Tam SC, Tien P, *et al.* Sifuvirtide, a potent HIV fusion inhibitor peptide. *Biochem Biophys Res Commun* 2009,**382**:540-544.
136. Eckert DM, Kim PS. Design of potent inhibitors of HIV-1 entry from the gp41 N-peptide region. *Proc Natl Acad Sci U S A* 2001,**98**:11187-11192.
137. Nishikawa H, Nakamura S, Kodama E, Ito S, Kajiwara K, Izumi K, *et al.* Electrostatically constrained alpha-helical peptide inhibits replication of HIV-1 resistant to enfuvirtide. *Int J Biochem Cell Biol* 2009,**41**:891-899.
138. Qi Z, Pan C, Lu H, Shui Y, Li L, Li X, *et al.* A recombinant mimetics of the HIV-1 gp41 prehairpin fusion intermediate fused with human IgG Fc fragment elicits neutralizing antibody response in the vaccinated mice. *Biochem Biophys Res Commun* 2010,**398**:506-512.
139. Eggink D, Bontjer I, Langedijk JP, Berkhout B, Sanders RW. Resistance of human immunodeficiency virus type 1 to a third-generation fusion inhibitor requires multiple mutations in gp41 and is accompanied by a dramatic loss of gp41 function. *J Virol* 2011,**85**:10785-10797.
140. Dwyer JJ, Wilson KL, Davison DK, Freel SA, Seedorff JE, Wring SA, *et al.* Design of helical, oligomeric HIV-1 fusion inhibitor peptides with potent activity against enfuvirtide-resistant virus. *Proc Natl Acad Sci U S A* 2007,**104**:12772-12777.
141. Yu H, Tudor D, Alfsen A, Labrosse B, Clavel F, Bomsel M. Peptide P5 (residues 628-683), comprising the entire membrane proximal region of HIV-1 gp41 and its calcium-binding site, is a potent inhibitor of HIV-1 infection. *Retrovirology* 2008,**5**:93.
142. Zhao L, O'Reilly MK, Shultz MD, Chmielewski J. Interfacial peptide inhibitors of HIV-1 integrase activity and dimerization. *Bioorg Med Chem Lett* 2003,**13**:1175-1177.
143. Kong R, Wang C, Ma X, Liu J, Chen W. Peptides design based on the interfacial helix of integrase dimer. *Conf Proc IEEE Eng Med Biol Soc* 2005,**5**:4743-4746.
144. Li HY, Zawahir Z, Song LD, Long YQ, Neamati N. Sequence-based design and discovery of peptide inhibitors of HIV-1 integrase: insight into the binding mode of the enzyme. *J Med Chem* 2006,**49**:4477-4486.
145. Levin A, Hayouka Z, Helfer M, Brack-Werner R, Friedler A, Loyter A. Stimulation of the HIV-1 integrase enzymatic activity and cDNA integration by a peptide derived from the integrase protein. *Biopolymers* 2010,**93**:740-751.
146. Maroun RG, Gayet S, Benleulmi MS, Porumb H, Zargarian L, Merad H, *et al.* Peptide inhibitors of HIV-1 integrase dissociate the enzyme oligomers. *Biochemistry* 2001,**40**:13840-13848.
147. Sourgen F, Maroun RG, Frere V, Bouziane M, Auclair C, Troalen F, *et al.* A synthetic peptide from the human immunodeficiency virus type-1 integrase exhibits coiled-coil properties and interferes with the in vitro integration activity of the enzyme. Correlated biochemical and spectroscopic results. *Eur J Biochem* 1996,**240**:765-773.
148. Maroun RG, Krebs D, El Antri S, Deroussent A, Lescot E, Troalen F, *et al.* Self-association and domains of interactions of an amphipathic helix peptide inhibitor of HIV-1 integrase assessed by analytical ultracentrifugation and NMR experiments in trifluoroethanol/H(2)O mixtures. *J Biol Chem* 1999,**274**:34174-34185.
149. Azzi S, Parissi V, Maroun RG, Eid P, Mauffret O, Fermandjian S. The HIV-1 integrase alpha4-helix involved in LTR-DNA recognition is also a highly antigenic peptide element. *PLoS One* 2010,**5**:e16001.
150. Krebs D, Maroun RG, Sourgen F, Troalen F, Davoust D, Fermandjian S. Helical and coiled-coil-forming properties of peptides derived from and inhibiting human immunodeficiency virus type 1 integrase assessed by 1H-NMR--use of NH temperature coefficients to probe coiled-coil structures. *Eur J Biochem* 1998,**253**:236-244.
151. Zhao L, Chmielewski J. Inhibition of HIV-1 integrase dimerization and activity with crosslinked interfacial peptides. *Bioorg Med Chem* 2012.
152. Niedrig M, Gelderblom HR, Pauli G, Marz J, Bickhard H, Wolf H, *et al.* Inhibition of infectious human immunodeficiency virus type 1 particle formation by Gag protein-derived peptides. *J Gen Virol* 1994,**75 (Pt 6)**:1469-1474.

153. Bocanegra R, Nevot M, Domenech R, Lopez I, Abian O, Rodriguez-Huete A, *et al.* Rationally designed interfacial peptides are efficient in vitro inhibitors of HIV-1 capsid assembly with antiviral activity. *PLoS One* 2011,**6**:e23877.
154. Hilpert K, Behlke J, Scholz C, Misselwitz R, Schneider-Mergener J, Hohne W. Interaction of the capsid protein p24 (HIV-1) with sequence-derived peptides: influence on p24 dimerization. *Virology* 1999,**254**:6-10.
155. Morris MC, Robert-Hebmann V, Chaloin L, Mery J, Heitz F, Devaux C, *et al.* A new potent HIV-1 reverse transcriptase inhibitor. A synthetic peptide derived from the interface subunit domains. *J Biol Chem* 1999,**274**:24941-24946.
156. Narumi T, Komoriya M, Hashimoto C, Wu H, Nomura W, Suzuki S, *et al.* Conjugation of cell-penetrating peptides leads to identification of anti-HIV peptides from matrix proteins. *Bioorg Med Chem* 2012,**20**:1468-1474.
157. Oz Gleenberg I, Avidan O, Goldgur Y, Herschhorn A, Hizi A. Peptides derived from the reverse transcriptase of human immunodeficiency virus type 1 as novel inhibitors of the viral integrase. *J Biol Chem* 2005,**280**:21987-21996.
158. Zawahir Z, Neamati N. Inhibition of HIV-1 integrase activity by synthetic peptides derived from the HIV-1 HXB2 Pol region of the viral genome. *Bioorg Med Chem Lett* 2006,**16**:5199-5202.
159. Oz Gleenberg I, Herschhorn A, Goldgur Y, Hizi A. Inhibition of human immunodeficiency virus type-1 reverse transcriptase by a novel peptide derived from the viral integrase. *Arch Biochem Biophys* 2007,**458**:202-212.
160. Louis JM, Dyda F, Nashed NT, Kimmel AR, Davies DR. Hydrophilic peptides derived from the transframe region of Gag-Pol inhibit the HIV-1 protease. *Biochemistry* 1998,**37**:2105-2110.
161. Davis DA, Brown CA, Singer KE, Wang V, Kaufman J, Stahl SJ, *et al.* Inhibition of HIV-1 replication by a peptide dimerization inhibitor of HIV-1 protease. *Antiviral Res* 2006,**72**:89-99.
162. Davis DA, Tebbs IR, Daniels SI, Stahl SJ, Kaufman JD, Wingfield P, *et al.* Analysis and characterization of dimerization inhibition of a multi-drug-resistant human immunodeficiency virus type 1 protease using a novel size-exclusion chromatographic approach. *Biochem J* 2009,**419**:497-506.
163. Liu F, Boross PI, Wang YF, Tozser J, Louis JM, Harrison RW, *et al.* Kinetic, stability, and structural changes in high-resolution crystal structures of HIV-1 protease with drug-resistant mutations L24I, I50V, and G73S. *J Mol Biol* 2005,**354**:789-800.
164. Veljkovic N, Branch DR, Metlas R, Prljic J, Manfredi R, Stringer WW, *et al.* Antibodies reactive with C-terminus of the second conserved region of HIV-1gp120 as possible prognostic marker and therapeutic agent for HIV disease. *J Clin Virol* 2004,**31 Suppl 1**:S39-44.
165. Franke R, Hirsch T, Overwin H, Eichler J. Synthetic mimetics of the CD4 binding site of HIV-1 gp120 for the design of immunogens. *Angew Chem Int Ed Engl* 2007,**46**:1253-1255.
166. Carlier E, Mabrouk K, Moulard M, Fajloun Z, Rochat H, De Waard M, *et al.* Ion channel activation by SPC3, a peptide derived from the HIV-1 gp120 V3 loop. *J Pept Res* 2000,**56**:427-437.
167. Sakaida H, Hori T, Yonezawa A, Sato A, Isaka Y, Yoshie O, *et al.* T-tropic human immunodeficiency virus type 1 (HIV-1)-derived V3 loop peptides directly bind to CXCR-4 and inhibit T-tropic HIV-1 infection. *J Virol* 1998,**72**:9763-9770.
168. Haynes BF, Ma B, Montefiori DC, Wrin T, Petropoulos CJ, Sutherland LL, *et al.* Analysis of HIV-1 subtype B third variable region peptide motifs for induction of neutralizing antibodies against HIV-1 primary isolates. *Virology* 2006,**345**:44-55.
169. Zolla-Pazner S, Cohen S, Pinter A, Krachmarov C, Wrin T, Wang S, *et al.* Cross-clade neutralizing antibodies against HIV-1 induced in rabbits by focusing the immune response on a neutralizing epitope. *Virology* 2009,**392**:82-93.
170. Chertov O, Zhang N, Chen X, Oppenheim JJ, Lubkowski J, McGrath C, *et al.* Novel peptides based on HIV-1 gp120 sequence with homology to chemokines inhibit HIV infection in cell culture. *PLoS One* 2011,**6**:e14474.
171. Marchio S, Alfano M, Primo L, Gramaglia D, Butini L, Gennero L, *et al.* Cell surface-associated Tat modulates HIV-1 infection and spreading through a specific interaction with gp120 viral envelope protein. *Blood* 2005,**105**:2802-2811.

172. De Houwer S, Demeulemeester J, Thys W, Taltynov O, Christ F, Debyser Z. Identification of residues in the C-terminal domain of HIV-1 integrase that mediate binding to TRN-SR2. *J Biol Chem* 2012.
173. Armon-Omer A, Graessmann A, Loyter A. A synthetic peptide bearing the HIV-1 integrase 161-173 amino acid residues mediates active nuclear import and binding to importin alpha: characterization of a functional nuclear localization signal. *J Mol Biol* 2004,**336**:1117-1128.
174. Xiao H, Neuveut C, Tiffany HL, Benkirane M, Rich EA, Murphy PM, *et al.* Selective CXCR4 antagonism by Tat: implications for in vivo expansion of coreceptor use by HIV-1. *Proc Natl Acad Sci U S A* 2000,**97**:11466-11471.
175. Ghezzi S, Noonan DM, Aluigi MG, Vallanti G, Cota M, Benelli R, *et al.* Inhibition of CXCR4-dependent HIV-1 infection by extracellular HIV-1 Tat. *Biochem Biophys Res Commun* 2000,**270**:992-996.
176. Jain S, Rosenthal KL. The gp41 epitope, QARVLAVERY, is highly conserved and a potent inducer of IgA that neutralizes HIV-1 and inhibits viral transcytosis. *Mucosal Immunol* 2011,**4**:539-553.
177. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* 2013,**30**:2725-2729.

Chapter 4

HIV-1 Gag C-terminal amino acid substitutions emerging under selective pressure of protease inhibitors

"All great things start from small beginnings."

— Marcus Tullius Cicero

This chapter is adapted from my article:
Guangdi Li, Jens Verheyen, Kristof Theys, Supinya Piampongsant, Kristel Van Laethem, Anne-Mieke Vandamme. HIV-1 Gag C-terminal amino acid substitutions emerging under selective pressure of protease inhibitors in patient populations infected with different HIV-1 subtypes. Retrovirology. 2014 (accepted).

I proposed the idea, performed the statistical analysis and drafted the manuscript. The improvement of the paper was supported with substantial help from Prof. Anne-Mieke Vandamme, Prof. Kristel Van Laethem and Dr. Kristof Theys, as well as advices and corrections from other coauthors. I sincerely thank Fossie Ferreira, Yoeri Schrooten, Lore Vinken, Jasper Edgar Neggers, Nádia Conceição Neto, Liana Eleni Kafetzopoulou, Dan Clements and Jurgen Vercauteren for technical assistance and valuable contributions to the analysis.

4.1 Summary

HIV-1 Gag amino acid substitutions associated with protease inhibitor (PI) treatment have mainly been reported in subtype B, while information on other subtypes is scarce. Using sequences from 11613 patients infected with different HIV-1 subtypes, we evaluated the prevalence of 93 Gag amino acid substitutions and their association with genotypic PI resistance. A significant association was found for 13 Gag substitutions, including A431V in both subtype B and CRF01_AE. K415R in subtype C and S451G in subtype B were newly identified. Most PI-associated Gag substitutions are located in the flexible C-terminal domain, revealing the key role this region plays in PI resistance.

4.2 Introduction

An amino acid substitution is commonly defined as an amino acid change between two consecutive sequences based on longitudinal data [1, 2]. Amino acid substitutions in HIV-1 protease, commonly called resistance mutations if they confer HIV-1 drug resistance, are known to emerge under selective pressure of protease inhibitors (PIs) [3]. As an alternative mechanism, HIV-1 can escape PI selective pressure by the selection of substitutions in the protease substrate Gag [1, 4-7]. Such Gag substitutions arising during PI-based treatment have mostly been characterized in HIV-1 subtype B (**Table 4.1**), while only a few studies have focused on non-B subtypes using small cohorts of patients (**Table 4.2**). Gag variability has been shown to impact PI susceptibility in a subtype-dependent manner [4, 6], warranting a comprehensive analysis of PI-associated Gag substitutions across different subtypes. Here, we identified novel Gag substitutions in HIV-1 non-B subtypes using longitudinal data from patients failing PI-based therapy. Moreover, we evaluated the prevalence of the newly identified and the previously reported Gag substitutions in different HIV-1 subtypes and investigated their association with genotypic PI resistance using a large sequence dataset.

Table 4.1: Summary of HIV-1 subtype B Gag substitutions observed during PI-based treatment.

Gag protein	Gag position	Gag substitutions *	Reference
Matrix	12	E12K	[8]
Matrix	62	G62R	[9]

Matrix	75	L75R	[8]
Matrix	76	R76K	[4]
Matrix	79	Y79F	[10]
Matrix	81	T81A	[10]
Matrix	125	S125K	[9]
Matrix	132	Y132F	[9]
Capsid	219	H219Q	[8, 11]
p2	370	V370A, V370I, V370M	[12]
p2	373	S373Q	[5, 13]
p2	374	A374G,A374N,A374P,A374S,A374T	[5, 14]
p2	375	T375N	[5]
Nucleocapsid	390	V390D	[8]
Nucleocapsid	409	R409K	[8]
Nucleocapsid	431	A431V	[7, 13, 15-23]
p1	435	G435E, G435R	[1, 17, 24]
p1	436	K436E, K436N, K436R	[14, 17, 18, 24, 25]
p1	437	I437A, I437T, I437V	[2, 14, 17, 18, 20, 23-26]
p1	438	W438R	[17]
p1	440	S440C	[24]
p6	449	L449F, L449P, L449Q, L449V	[2, 13-15, 18, 20, 23, 24, 27-32]
p6	451	S451I, S451N, S451T	[32, 33]
p6	452	R452K, R452S	[15, 18, 23, 34]
p6	453	P453L, P453T	[14, 16, 18, 20, 21, 23, 28, 29, 31, 34]
p6	459	P459I	[35]

*: Only Gag substitutions associated with FDA-approved PIs in HIV-1 subtype B are summarized. The substitution is expressed relative to the subtype B consensus sequence (http://www.hiv.lanl.gov/).

Table 4.2: Summary of Gag amino acid substitutions in HIV-1 non-B subtypes observed during PI-based treatment.

Gag amino acid substitutions in non-B subtypes *	Number of patients infected by non-B subtypes	Reference
A431V	A or F # (n=4)	[21]
K436R, **N451S**	C (n=1)	[21]
L363F, A364G, **A374T**, I376V, M378V, R380K, **K436E**, G443R	G (n=2), 01_AE (n=1), 02_AG (n=4)	[14]
V128A/T/I, Q130R, **Y132F**, V135M, V362I, A373T, **A374T**, A375T, I376A/V/M, K380R, S381G, N382K, E428D/Q, Q430R/G/V/I, **A431V, K436R**, L449I, **N451S**, **R452K**, P453I	G (n=21)	[36]
N375S, G381R	A1 (n=2)	[2]
G381S, G446E	02_AG (n=1)	[2]
V135I, I376V, L486F	01_AE (n=1)	[2]

P453L/T/I	F # (n=61)	[37]
M138L, F363L, L363W, **A374T**, V374A, R387K, N389T, K411Q, K415R, G420A, P422Q, T427P, P445L, S451G, R452G, **P453L**, P453Ins, I469T, P472S, P474L, E477Q	A1 (n=1), C (n=6), D (n=1), F1 (n=1), J (n=1), 01_AE (n=1), 02_AG (n=1)	Our study

*: Non-B Gag substitutions reported during PI-based treatment either in the literature or in our study. The substitutions also identified in subtype B are indicated in bold.

#: Information of HIV-1 subtype or sub-subtype was ambiguous or not available.

4.3 Findings

We first investigated the emergence of non-B Gag substitutions during PI-based treatment in a cohort of 1068 patients followed at the University Hospital of Leuven, for which virological outcome and treatment information were available [38]. Our protocol and quality control of viral sequencing and viral load tests have been described previously [39, 40]. The sequences with associated information are available through Euresist (http://www.euresist.org). For 69 patients infected with HIV-1 non-B subtypes and receiving PI-based treatment for at least three months, sequence information for Gag, protease and reverse transcriptase (RT) was available at baseline and at treatment failure, which was defined according to the guidelines of the European AIDS Clinical Society (EACS) (http://www.eacsociety.org/). Under drug selective pressure, 21 different substitutions at 18 Gag positions were identified among 12 patients, of whom 11 harbored Gag substitutions in the presence of (pre-existing or simultaneously acquired) drug resistance mutations in protease or RT (**Figure 4.1**, **Table 4.3**). Gag substitution P453Ins (insertion: EPTAPP) emerged in patient 343 in the absence of PI and RTI resistance mutations. Some substitutions were from a less to a more common amino acid such as M138L. Specifically, patients failing LPV/r-based regimens developed one of the following Gag substitution patterns: L363W+E477Q, F363L+N389T+P422Q+P455L, K411Q, P472S+P474L, K415R+I469T, M138L, A374T or G420A. Patients failing DRV/r-based regimens developed Gag substitution patterns P453Ins or T427P+R452G. Patients failing an ATV/r-based regimen developed Gag substitution patterns: P453L or V374A+R387K+S451G+P453Ins. A patient failing a regimen containing FPV/r and SQV/r developed L363W. Longitudinal data from 34 PI-naïve patients infected with non-B subtypes revealed the emergence of one Gag substitution (V370A) in a single patient. Overall, when analyzing all subtypes, the proportion of PI-treated patients

with Gag substitutions was much higher than that of PI-naïve patients (17.4% (12/69) vs 2.9% (1/34), p-value = 0.037).

Table 4.3: Summary of Gag, protease and RT substitutions in 12 patients of the Leuven cohort.

Patient ID	Subtype	Sampling day	Gag substitution	Protease variants and PI resistance mutations	RT variants and RTI resistance mutations
123	A1	2002-10-08	363L+477E	10V	49R+103N+108I+118V+173[S,L]+179I+184V +189[I,V]+215[N,T,Y,S]+225[H,P]+238[K,T] +248T+303L +348[N,I]
		2003-12-10	363L+477Q	10V+20T	49[K,R]+103N+108I+118V+173L+179L+184M +189I+215Y+225P+238T+248T+303[W,L]+348na
		2004-10-27	363W+477Q	10V+20T	49K+103N+108I+118[I,V]+173L+179L+184M +189I+215Y+225P+ 238T+248[T,I]+303L+348na
343	C	2010-03-17	453P	64I+70[K,R]	40[E,D]+49[K,R]+106[M,V]+123[S,G]+126[K,R] +165[T,I]+175N+277[K,R]+281R
		2012-11-30	453PTAPPE	64[I,L]+70K	40E+49K+106V+123S+126K+165I+175[N,Y] +277K+281[K,R]
357	C	2004-07-26	363F+389N +422P+445P	74S	13[R,K]+86D+148V+154[R,K]+280[Y,C]
		2005-09-05	363L+389T +422Q+445L	74S	13K+86[D,G]+148[V,I]+154K+280C
445	C	2007-06-27	411K	60[D,E]+66I	3S+28K+32E+36A+49K+60I+65K+67D+70K +75[I,V]+101K+103N+123[N,S]+173V+177D +184M+190G+207[E,G]+211R+214L+215[T,I]+ 219K+277[K,R]+ 278Q+281R
		2008-01-02	411K	60E+66I	3S+28K+32E+36A+49K+60I+65R+67D+70K +75V+101K+103N+123S+173V+177D+184M +190G+207E+211R+214L+215[T,I]+219K +277R+278[Q,R]+281K
		2011-07-15	411Q	60D+66[I,M]	3S+28E+32K+36E+49[K,R]+60V+65K+67N +70R+75V+101K+103K+123S+173A+177E +184V+190A+207E+211K+214[L,F]+215T +219K+277R+278Q+281K
		2012-04-04	411Q	60D+66I	3[N,S,D,G]+28[K,E]+32K+36E+49K+60V+65K +67N+70R+75V+101[K,E]+103K+123S+173A +177E+184V+190A+207E+211K+214[L,F] +215T+219[K,E]+277R+278Q+281K
834	C	2003-02-12	472P+474P	20R+62V	103N+173T +184V
		2007-01-04	472S+474L	20R+62V	103K+173A +184M
1039	C	2010-08-12	427T +452R	62[I,V]+72I+ 74S+82[I,V]	8[I,V]+28E+39E+48[T,S]+53[E,D]+67D +121[H,D]+123[N,S,D,G]+135I+142[I,V] +166[K,R]+184M+208H+214F+241V +286[T,A]+324[E,D] +334Q
		2011-02-14	427T +452R	62I+72I+ 74S+82V	8V+28E+39E+48T+53E+67[N,D]+121D+123S +135I+142[I,V]+166[K,R]+184[I,M]+208H +214F+241V+286A+324E+334[Q,H]
		2011-04-01	427T +452R	62[I,V]+72I+ 74S+82[I,V]	8V+28E+39E+48[T,S]+53[E,D]+67[N,D]+121D +123S+135I+142I+166[K,R]+184V+208[H,Y] +214F+241V+286A+324E+334na
		2011-08-08	427T +452G	62V+72I+74S+82I	8V+28K+39E+48T+53E+67D+121D+123S+135L +142I+166K+184V+208H+214F+241V+286A +324E+334Q
		2013-04-22	427P+452G	62[I,V]+72[I,M]	8V+28E+39[K,E]+48T+53E+67D+121D+123S

				+74S+ 82[I,V]	+135[T,I,S,L]+142I+166K+184V+208H +214[L,F]+241[V,L]+286A+324E+334Q
1075	C	2005-11-07	363L	10I+18H+33F+ 43T+48V+54A+ 62V+74S+82A+89I	35T+41L+44D+67N+84[T,S]+101K+106M+ 118I+184V+190A+210W+215Y+219N+227L
		2006-03-01	363W	10I+18H+33F+43T +48V+54A+62V +74S+82A+89I	35T +41L+44D+67N+84T+101[K,E] +106[I,M,V]+118I+184[M,V]+190A+ 210W+215Y+219N+227L
		2006-05-15	363L	10I+18[Q,H]+33F +43T+48V+54A+ 62V+74S+82A+89I	35[T,I]+41L+44D+67N+84T+101[K,E] +106V+118I+184[M,V]+190A+210W +215Y+219N+227L
552	D	2002-02-18	415K+469I	13I+14K+62I +72I+93[I,L]	20[K,R]+136N+275K+294P+297A
		2003-07-28	415K+469T	13I+14K+62I +72I+93L	20[K,R]+136N+275[K,R]+294P+ 297A+334R+335G
		2004-03-22	415K+469T	13I+14K+62[I,V] +72[I,V]+93L	20[K,R]+136N+275K+294P+297A +334R+335E
		2009-02-11	415K+469T	13[I,V]+14[K,R] +62I+72[I,V] +93[I,L]	20K+136N+275K+294T+297[T,A]+334R+335E
		2011-01-24	415R+469T	13[I,V]+14[K,R] +62[I,V]+72[I,V] +93[I,L]	20K+136[N,H]+275K+294na+297na+ 334na+335na
27	F1	2008-07-24	453P	10V+62V+69Y	207E+276V
		2010-07-14	453L	10V+62V+69Y	207K+276[I,V]
		2011-01-12	453L	10V+62V+69[H,Y]	207[K,E]+276V
666	J	2001-06-13	138M+374A	10I+20R+35E+54I +60[N,S]+62V+ 82V+89M	67D+70K+101K+174[K,R]+184M+200I+288A +294T+322T
		2003-11-12	138L+374A	10I+20R+35E+54I+ 60S +62V+82V+89M	67D+70K+101K+174[K,R]+184V+200[T,I,A,V] +288[A,S]+294T+322T
		2004-11-03	374T	10I+20R+35D+54V+ 60S+62V+82A+89L	67N+70R+101E+174K+184M+200I+288A+294P +322A
407	01_AE	2009-11-02	374V+387R +451S+453P	79P	11K+184I+200A+281R+286A+ 304[E,A]+325L+326I
		2010-03-22	374A+387K +451G+ 453EPTAPP	79H	11[K,T]+184[I,M]+200[A,V]+281K+286[T,A] +304A+325I+326V
652	02_AG	2008-01-31	420G	20I	11R+27[T,S]+35I+106V+207G+276I+291D+292I +294T+311R
		2008-08-19	420A	20I	11[K,Q,R]+27T+35[T,I]+106[I,V]+207[E,G] +276[I,V]+291[E,D]+292[I,V]+294[T,P]+311[K,R]

For the first sequence, PI/RTI drug resistance mutations, detected by the drug resistance interpretation algorithms HIVdb v7.0 [41] and/or Rega V9.1 [42], are colored red. The other protease and RT variants are indicated in black. Ambiguous nucleotide letters are decomposed and translated into amino acids, which are indicated by brackets. Gag substitutions and PI/RTI resistance mutations with therapy changes are mapped in **Figure 4.1**. "na" indicates that the sequence does not cover the corresponding position.

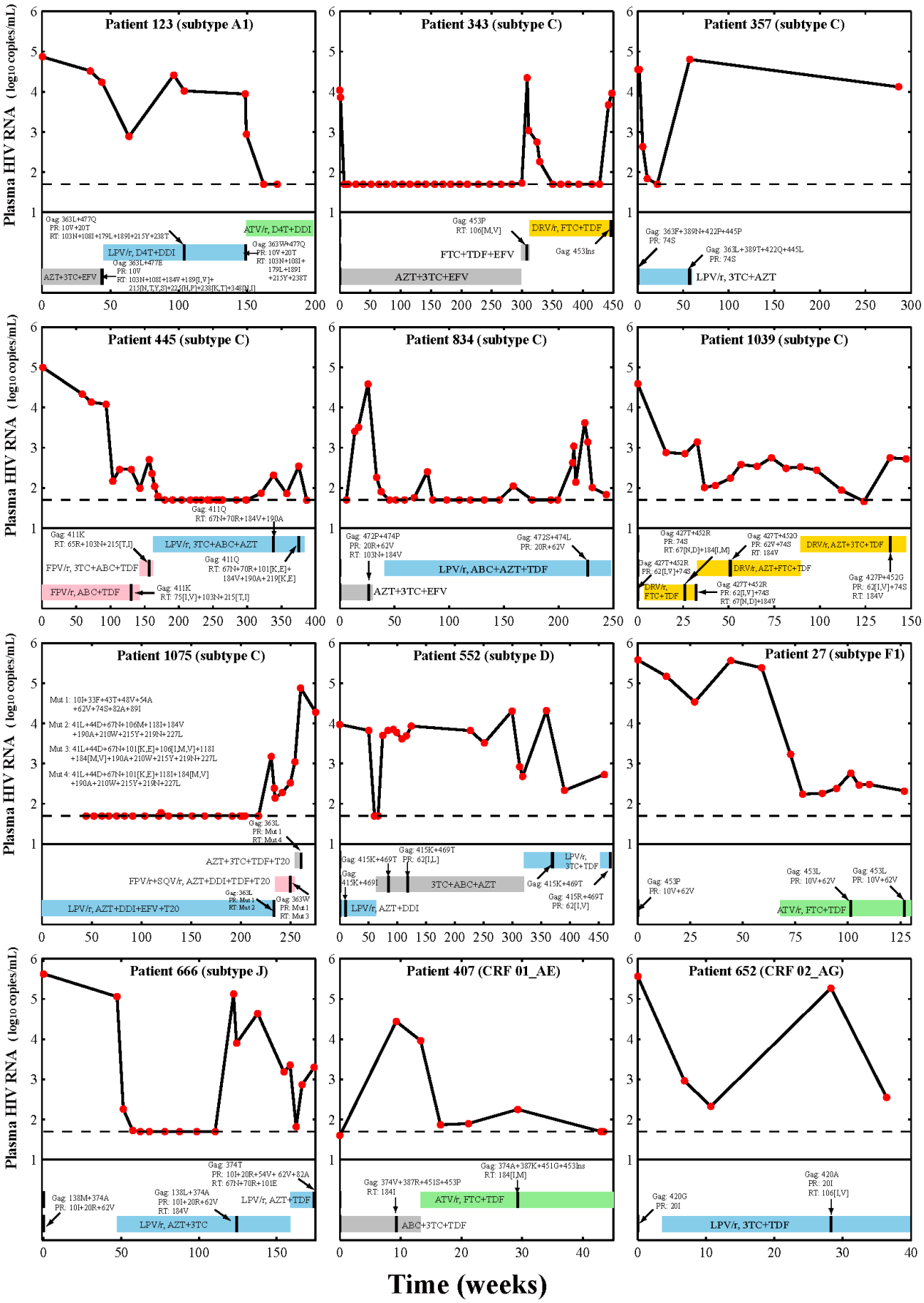

Patient 123 (subtype A1)
Patient 343 (subtype C)
Patient 357 (subtype C)
Patient 445 (subtype C)
Patient 834 (subtype C)
Patient 1039 (subtype C)
Patient 1075 (subtype C)
Patient 552 (subtype D)
Patient 27 (subtype F1)
Patient 666 (subtype J)
Patient 407 (CRF 01_AE)
Patient 652 (CRF 02_AG)
Plasma HIV RNA (log10 copies/mL)
Time (weeks)

Figure 4.1: Gag substitutions and PI or RTI resistance mutations in 12 patients from the Leuven cohort. Each subplot shows the data of one patient regarding the viral load, the treatment period and the emerging Gag substitutions and the PI/RTI resistance mutations. X- and Y-axes indicate the time (weeks) and the level of plasma HIV RNA (log_{10} copies/mL), respectively. For each subplot, red dots indicate the level of viral load and the dash line indicates the viral load cutoff at 50 copies per mL. Beneath the viral load plot, each treatment period is annotated by a colored bar with vertical black lines indicating the sequence sampling time. The blue, pink, green and yellow bars show PI-based treatments containing LPV/r, FPV/r, ATV/r and DRV/r, respectively. The grey bar indicates treatments lacking PIs. Multiple substitutions or mutations are shown using the plus symbol "+". Amino acids translated from ambiguous nucleotide letters are indicated by brackets. For patient 343, the insertion EPTAPP at position P453 is annotated as P453Ins. For patient 1075, the sets of PI or RTI resistance mutation are abbreviated (Mut 1-4) and listed in the subplot. **Table 4.3** provides the full list of Gag, protease and RT substitutions in these 12 patients.

For our second analysis, we compiled a comprehensive list of 93 Gag substitutions at 55 positions in B and non-B subtypes observed in PI-treated patients, based on literature results or our first analysis as described above (**Table 4.1, Table 4.2**). Next, we systematically evaluated the prevalence of these variants in major HIV-1 subtypes using 10865 full-length Gag sequences retrieved from the HIV Los Alamos database (one sequence per patient) (**Table 4.4**). Sequence alignment and quality control have been described previously [43]. We found that the prevalence of 62 (66.7%) Gag variants at 39 positions was above 1% in at least one subtype or CRF (A1, B, C, D, F1, G, CRF01_AE, CRF02_AG). Among the 55 Gag positions, only 363 and 455 were highly conserved with less than 1% overall amino acid variation in every subtype and CRF in our dataset (**Figure 4.2**A). Moreover, 77 of these 93 variants (82.8%) were found at 42 positions located in the Gag C-terminal domain (positions: 362-500).

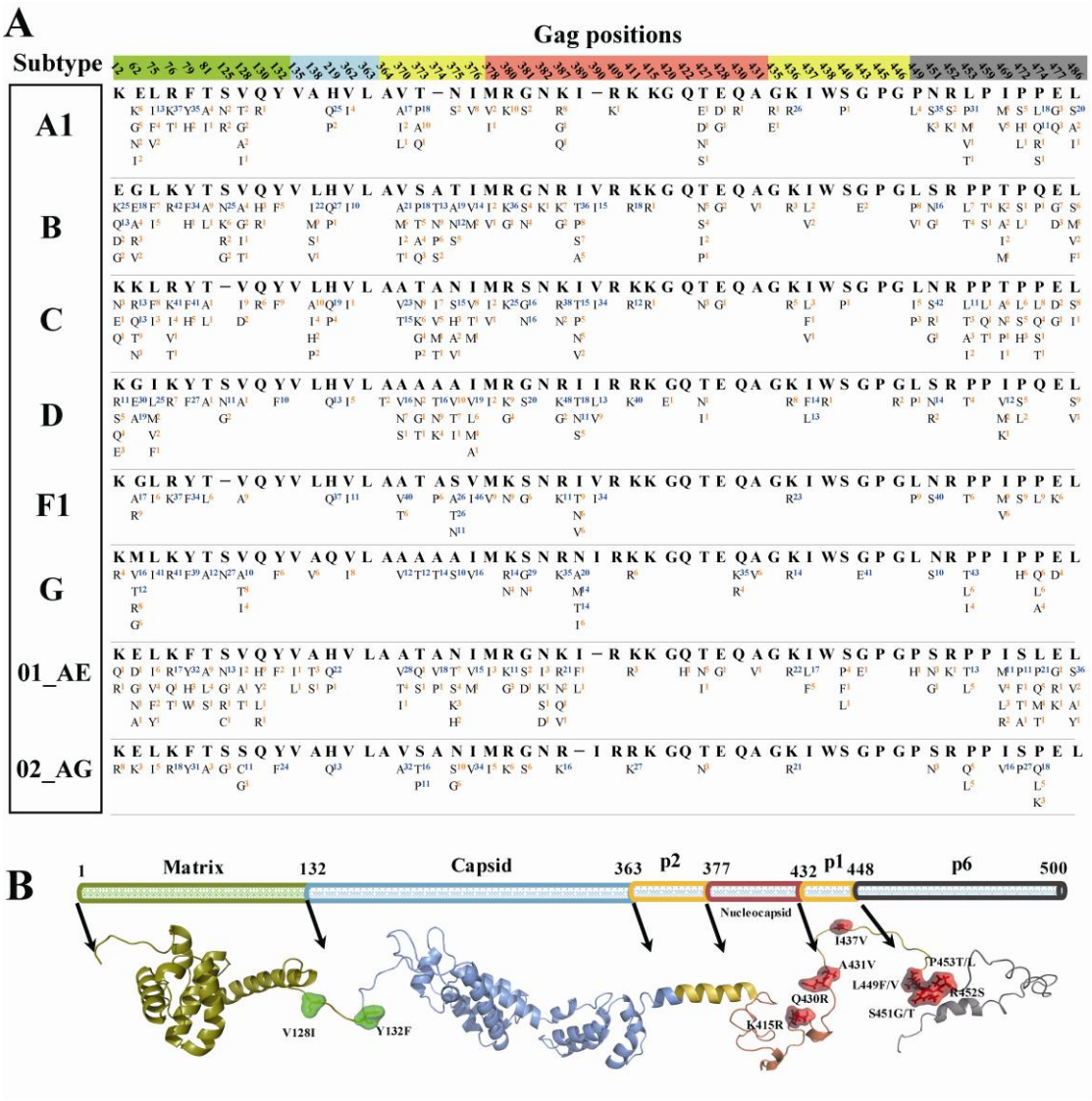


Figure 4.2: Prevalence of Gag amino acid variants reported in patients failing PI-based therapies and their mapping to HIV-1 protein structures.

(A) Prevalence of amino acid variations at 55 Gag positions in 8 HIV-1 subtypes (A1, B, C, D, F1, G, 01_AE and 02_AG) given the Los Alamos full-length Gag sequence dataset (Table 2). Only Gag positions where amino acid substitutions have been observed during PI-based treatment are shown. For each position, the HXB2 index is shown at the top, followed by the most prevalent amino acids (bold) and amino acid variations in our sequence datasets. Amino acids with blue superscripts have prevalence above 10% and other amino acids have orange superscripts.

(B) Structural representation of Gag polyprotein and mapping of the 13 PI-associated Gag substitutions identified in Table 3. The annotation of Gag polyproteins is shown at the top. Individual Gag protein structures are shown at the bottom. Gag substitutions are annotated and colored accordingly. Red surfaces indicate PI-associated Gag substitutions at the Gag C-terminal domain; other substitutions are shown in green. PDB data of Gag protein structures: matrix, 1HIW; capsid, 3NTE; p2, 1U57; nucleocapsid, 2M3Z; p6, 2C55. Visualization software: PyMOL V1.5 (http://www.pymol.org/).

As treatment information of the 10865 full-length gag nucleotide sequences was largely lacking, our third analysis aimed to evaluate whether these 93 Gag variants were significantly associated with genotypic PI resistance. Among the 11613 sequences pooled from the Leuven and the Los Alamos datasets (**Table 4.4**), 6645 spanned both the gag and the full-length protease regions, and were translated into amino acid sequences for our analysis. Using the drug resistance interpretation algorithms HIVdb V7.0 [41] and Rega V9.1 [42], 660 sequences were concordantly estimated to be partially or fully resistant to at least one PI, and 5657 sequences were concordantly estimated to be fully susceptible to all PIs (**Table 4.5**). Sequences with discordant estimates of PI susceptibility were excluded from our analysis. Fisher's exact tests were then used to compare the amino acid prevalence between these PI-susceptible and PI-resistant datasets. Of the 93 Gag variants, 16 at 13 amino acid positions were associated with (partial or full) PI resistance in at least one HIV-1 subtype (p-value < 0.05, **Table S4.1**). After multiple testing correction using the false discovery rate approach described in [44], 13 Gag variants at 10 positions remained significantly PI-associated within individual subtypes (adjusted p-value < 0.05), including 11 variants located in the Gag C-terminal domain (**Figure 4.2**B, **Table 4.6**). Our analysis successfully identified the known PI-associated Gag substitution A431V, strengthening the validity of our approach. As the only PI-associated Gag substitution found in more than one subtype, A431V had a high prevalence in the PI-resistant strains of subtype B (13.5%) and CRF01_AE (18.2%) (**Table 4.6**). Interestingly, of the 21 Gag substitutions observed in our first analysis, K415R and S451G were newly identified to be significantly associated with genotypic PI resistance in subtypes C and B respectively, suggesting a possible involvement in PI-resistance.

Table 4.4: Summary of Leuven and Los Alamos sequence datasets.

Subtype	Los Alamos dataset	Leuven dataset			Total number of patient
	Number of Gag sequence *	Number of Gag sequence	Number of PI-naïve patient	Number of PI-treated patient	
A1	1648	167	72	19	1739
B	4131	639	313	57	4501
C	2780	198	58	24	2862
D	443	42	20	9	472
F1	35	38	25	4	64
G	49	1	1	0	50
J	3	8	1	2	6

01_AE	1714	72	45	5	1764
02_AG	62	139	71	22	155
Total	10865	1304	606	142	11613

*: Number of Gag sequences in different HIV-1 subtypes (one sequence per patient).

Table 4.5: Summary of PI-resistant and PI-susceptible sequence datasets

Subtype	Number of PI-susceptible gag-protease sequences *	Number of PI-resistant gag-protease sequences #	Total
A	185+72=257	6+3=9	266
B	1820+313=2133	434+31=465	2598
C	1728+58=1786	119+18=137	1923
D	98+20=118	1+0=1	119
F	21+25=46	3+0=3	49
G	33+0=33	14+0=14	47
01_AE	1112+45=1157	22+2=24	1181
02_AG	55+71=126	3+4=7	133
Total	5657	660	6317

*: Number of PI-susceptible Gag-protease sequences used in this study. These sequences were amino acid sequences translated from nucleotide sequences obtained from the Los Alamos + Leuven datasets. PI-susceptible sequences were estimated to be fully susceptible to all PIs by both the HIVdb v7.0 [41] and the Rega V9.1 [42] algorithms.

#: Number of PI-resistant Gag-protease sequences used in this study. These sequences were amino acid sequences translated from nucleotide sequences obtained from the Los Alamos + Leuven datasets. PI-resistant sequences were estimated to be partially or fully resistant to at least one PI by both the HIVdb v7.0 [41] and the Rega V9.1 [42] algorithms. All the Los Alamos sequences encode the full-length Gag polyprotein.

Table 4.6: Prevalence of PI-associated Gag substitutions in individual HIV-1 subtypes.

Gag substitutions *	Subtype	Prevalence #		P-value	Adjusted p-value
		PI-resistant dataset	PI-susceptible dataset		
V128I	B	5.8%(7/121[&])	0.9%(6/638)	0.002	0.024
Y132F	B	10.7%(13/122)	3.4%(22/639)	0.004	0.035
K415R	C	2.5%(3/119)	0.0%(0/1727)	<0.0001	0.012
Q430R	C	2.5%(3/119)	0.1%(1/1727)	0.003	0.046
A431V	B	13.5%(23/170)	0.1%(1/787)	<0.0001	<0.0001
	01_AE	18.2%(4/22)	0.7%(8/1111)	<0.0001	0.007
I437V	B	8.9%(15/168)	1.7%(13/784)	<0.0001	<0.0001
L449F	B	5.6%(21/377)	0.5%(7/1352)	<0.0001	<0.0001
L449V	B	4.8%(18/377)	0.9%(12/1352)	<0.0001	<0.0001
S451G	B	3.4%(13/378)	1.3%(17/1348)	0.008	0.041
S451T	B	2.1%(8/378)	0.0%(0/1348)	<0.0001	<0.0001
R452S	B	3.4%(13/384)	0.3%(4/1374)	<0.0001	<0.0001
P453T	C	21.8%(26/119)	3.1%(53/1722)	<0.0001	<0.0001
P453L	B	18.5%(71/384)	7.1%(99/1399)	<0.0001	<0.0001

*: A list of Gag substitutions whose prevalence differs significantly between sequences estimated to be (fully or partially) PI-resistant and sequences estimated to be PI-susceptible (see full reports in **Table S4.1**). One-tailed Fisher's exact tests were performed, and p-values were adjusted using multiple testing correction via the false discovery rate (FDR) approach [44].

#: Statistical analyses were only performed on individual subtype (B, C, G, 01_AE) datasets, which contained more than 10 (partially or fully) PI-resistant sequences. **Table 4.5** summarizes the subtype distribution of PI-resistant and PI-susceptible sequence datasets.

&: The numerator indicates the number of sequences for which the corresponding Gag position is covered; the denominator indicates the number of sequences displaying the respective amino acid substitutions.

4.4 Discussion and conclusions

To our knowledge, this study presents the first large-scale sequence analysis to establish statistical significance of PI-associated Gag substitutions in HIV-1 non-B

subtypes. Our longitudinal analysis of a clinical cohort of patients failing PI-based therapy confirmed that PI-treated patients developed more Gag substitutions than PI-naïve patients. The majority of these Gag substitutions emerged in the context of pre-existing or simultaneously acquired PI or RTI resistance mutations, confirming the important role of the known resistance mutations, while in some patients Gag substitutions emerged in the absence of resistance mutations (**Figure 4.1**, **Table 4.3**). Such Gag substitutions may therefore contribute to the virological failure of PI-based treatments. Based on two widely used genotypic interpretation algorithms, our comparative analysis found that only 13 (13.8%) of the 93 Gag substitutions emerging under PI selective pressure were significantly associated with genotypic PI resistance (**Table 4.6**). Particularly, the novel Gag substitutions K415R and S451G were identified in both our longitudinal and cross-sectional sequence analyses. This suggests that they may play a role in viral escape from PI selective pressure, partially contributing to the observed virological failure. Since virological outcome and treatment information is lacking for most sequences extracted from the HIV Los Alamos database, this limits our analysis to address the clinical impact of the newly identified substitutions with large-scale data.

Using small cohorts, previous studies suggested that different subtypes may develop different Gag substitutions [6, 45, 46]. We confirmed this hypothesis since only 9 of the 58 Gag substitutions reported in non-B subtypes (**Table 4.2**) were also observed in subtype B (**Table 4.1**). Among non-B Gag substitutions, 4 were significantly associated with genotypic PI resistance, of which only A431V was PI-associated in subtype B as well (**Table 4.6**). However, further evaluations on subtypes A2, D, F2, J, K and other CRFs are still needed due to the restriction of our study to particular subtypes. Interestingly, a predominant presence of PI-associated Gag substitutions at the flexible C-terminal domain of Gag (**Figure 4.2**B) leads us to suggest the hypothesis that PI-associated Gag substitutions tend to emerge in the structural flexible regions. These Gag substitutions can emerge along with protease drug resistance mutations as shown in our longitudinal sequence analysis (**Figure 4.1, Table 4.3**) and previous studies [15, 18]. Future studies are still needed to investigate the significance of coevolution between Gag substitutions and protease resistance mutations.

Overall, our findings showed different PI-associated substitutions in the Gag C-terminal domain across different subtypes, providing a roadmap to elucidate the role of Gag amino acid substitutions in the development of PI resistance.

4.5 Additional table

Table S4.1: Prevalence in individual HIV-1 subtypes of Gag amino acid variants observed during PI therapy.

Gag amino acid variant	Subtype *	Amino acid prevalence		p-value #	Adjusted p-value
		PI-resistant dataset	PI-susceptible dataset		
12E	C	0.8%(1/119)	0.2%(4/1724)	0.286	1
12K	B	25.4%(31/122)	25.3%(158/624)	0.923	1
62G	C	1.7%(2/117)	0.2%(4/1700)	0.054	0.741
62R	B	4.1%(5/121)	4.1%(26/639)	0.599	1
76K	C	48.3%(57/118)	44.7%(769/1719)	0.999	1
76R	01_AE	9.1%(2/22)	19.1%(212/1112)	0.958	1
79F	G	14.3%(2/14)	48.5%(16/33)	0.998	1
79Y	01_AE	9.1%(2/22)	33.7%(374/1111)	0.999	1
81A	B	15.6%(19/122)	7.5%(48/639)	0.021	0.165
81A	01_AE	4.5%(1/22)	10.9%(121/1112)	0.929	1
125K	B	4.1%(5/122)	1.6%(10/636)	0.087	0.559
128A	B	0.8%(1/121)	2.8%(18/638)	0.966	1
128A	G	15.4%(2/13)	9.7%(3/31)	0.532	1
128I	B	5.8%(7/121)	0.9%(6/638)	0.002	0.024
128I	C	7.7%(9/117)	11.8%(201/1709)	0.964	1
128I	G	15.4%(2/13)	0.0%(0/31)	0.101	1
128I	01_AE	4.5%(1/22)	0.7%(8/1089)	0.172	1
130R	B	2.5%(3/122)	1.3%(8/639)	0.263	0.925
130R	G	7.1%(1/14)	0.0%(0/33)	0.313	1
132F	B	10.7%(13/122)	3.4%(22/639)	0.004	0.035
132F	G	21.4%(3/14)	0.0%(0/33)	0.035	0.925
135I	B	1.6%(2/122)	0.3%(2/657)	0.121	0.625
135I	C	0.8%(1/119)	0.2%(3/1728)	0.236	1
135M	B	0.8%(1/122)	0.0%(0/657)	0.158	0.694
135M	G	7.1%(1/14)	0.0%(0/32)	0.319	1
138M	B	10.7%(13/122)	8.4%(55/657)	0.369	1
138M	C	0.8%(1/119)	0.2%(3/1721)	0.236	1
219Q	B	28.1%(34/121)	21.2%(141/665)	0.458	1
219Q	C	9.2%(11/119)	18.3%(314/1715)	0.999	1
219Q	01_AE	22.7%(5/22)	22.8%(253/1112)	0.768	1
362I	B	17.7%(22/124)	21.9%(149/680)	0.978	1
362I	C	2.5%(3/119)	1.6%(27/1727)	0.317	1
362I	G	21.4%(3/14)	0.0%(0/33)	0.035	0.925
363F	B	0.8%(1/124)	0.0%(0/680)	0.155	0.694
364G	C	0.8%(1/119)	0.1%(1/1728)	0.126	1
370A	B	17.8%(23/129)	14.7%(106/720)	0.491	1
370I	B	3.9%(5/129)	2.4%(17/720)	0.254	0.925
370M	B	4.7%(6/129)	6.1%(44/720)	0.83	1

370V	01_AE	36.4%(8/22)	31.3%(348/1111)	0.768	1
373A	B	4.9%(7/143)	4.8%(37/766)	0.6	1
373Q	B	0.7%(1/143)	2.0%(15/766)	0.938	1
373S	C	5.9%(2/34)	2.4%(27/1114)	0.229	1
373T	B	3.5%(5/143)	4.0%(31/766)	0.719	1
374G	B	2.8%(4/142)	1.1%(8/756)	0.114	0.625
374G	C	1.5%(1/65)	0.2%(1/591)	0.191	1
374N	B	8.5%(12/142)	5.4%(41/756)	0.168	0.706
374P	B	2.8%(4/142)	5.2%(39/756)	0.938	1
374S	B	1.4%(2/142)	3.0%(23/756)	0.929	1
374T	B	13.4%(19/142)	14.2%(107/756)	0.819	1
374V	B	0.7%(1/142)	2.5%(19/756)	0.97	1
374V	C	9.2%(6/65)	14.7%(87/591)	0.955	1
375A	B	28.0%(40/143)	17.2%(131/761)	0.089	0.559
375A	C	1.8%(2/111)	2.2%(34/1513)	0.727	1
375N	B	15.4%(22/143)	19.2%(146/761)	0.973	1
375N	B	15.4%(22/143)	19.2%(146/761)	0.973	1
375S	B	6.3%(9/143)	11.4%(87/761)	0.989	1
375T	C	0.9%(1/111)	0.5%(7/1513)	0.436	1
375T	G	14.3%(2/14)	0.0%(0/22)	0.171	1
376A	B	0.7%(1/144)	0.3%(2/763)	0.407	1
376A	G	15.4%(2/13)	3.0%(1/33)	0.227	1
376M	B	3.5%(5/144)	2.1%(16/763)	0.249	0.925
376M	C	2.5%(3/118)	0.3%(5/1724)	0.012	0.251
376M	G	7.7%(1/13)	3.0%(1/33)	0.512	1
376V	B	15.3%(22/144)	18.0%(137/763)	0.945	1
376V	C	9.3%(11/118)	6.8%(118/1724)	0.282	1
376V	G	23.1%(3/13)	12.1%(4/33)	0.412	1
378V	B	1.3%(2/153)	1.0%(8/769)	0.52	1
378V	C	0.8%(1/119)	0.9%(15/1726)	0.66	1
380K	B	29.4%(45/153)	25.0%(192/769)	0.771	1
380K	01_AE	4.5%(1/22)	10.5%(117/1111)	0.922	1
380R	G	7.1%(1/14)	15.2%(5/33)	0.91	1
381G	C	22.7%(27/119)	18.1%(310/1714)	0.488	1
381G	G	28.6%(4/14)	24.2%(8/33)	0.688	1
381S	B	2.6%(4/153)	2.3%(18/770)	0.527	1
382K	B	1.3%(2/153)	1.2%(9/768)	0.575	1
382K	G	7.1%(1/14)	0.0%(0/33)	0.313	1
387K	B	4.6%(7/152)	5.5%(42/769)	0.764	1
387R	01_AE	27.3%(6/22)	20.5%(223/1090)	0.526	1
389N	B	2.7%(4/150)	4.2%(32/758)	0.884	1
389N	C	4.4%(5/114)	2.1%(35/1682)	0.12	1
389T	C	9.6%(11/114)	17.8%(300/1682)	0.998	1
389T	G	8.3%(1/12)	0.0%(0/2)	0.867	1
415R	B	0.6%(1/169)	1.7%(13/786)	0.937	1
415R	C	2.5%(3/119)	0.0%(0/1727)	<0.0001	0.012
427P	B	1.2%(2/168)	0.6%(5/784)	0.363	1
427P	C	0.8%(1/119)	0.0%(0/1725)	0.065	0.771
430R	C	2.5%(3/119)	0.1%(1/1727)	0.003	0.046
430R	G	15.4%(2/13)	0.0%(0/33)	0.093	1
431V	B	13.5%(23/170)	0.1%(1/787)	<0.0001	<0.0001
431V	C	1.7%(2/119)	0.2%(4/1728)	0.054	0.741
431V	G	23.1%(3/13)	0.0%(0/33)	0.03	0.925

431V	01_AE	18.2%(4/22)	0.7%(8/1111)	<0.0001	0.007
435E	B	0.6%(1/170)	0.0%(0/785)	0.179	0.715
436R	B	4.7%(8/170)	4.8%(38/787)	0.637	1
436R	C	3.4%(4/119)	4.1%(71/1724)	0.75	1
436R	G	15.4%(2/13)	15.2%(5/33)	0.716	1
436R	01_AE	31.8%(7/22)	27.0%(300/1112)	0.702	1
437V	B	8.9%(15/168)	1.7%(13/784)	<0.0001	<0.0001
437V	C	1.7%(2/119)	0.6%(11/1726)	0.208	1
438R	B	0.6%(1/170)	0.3%(2/787)	0.446	1
449F	B	5.6%(21/377)	0.5%(7/1352)	<0.0001	<0.0001
449I	B	1.9%(7/377)	1.0%(13/1352)	0.132	0.644
449P	B	6.9%(26/377)	8.0%(108/1352)	0.871	1
449P	C	2.5%(3/119)	3.0%(52/1727)	0.716	1
449V	B	4.8%(18/377)	0.9%(12/1352)	<0.0001	<0.0001
451G	B	3.4%(13/378)	1.3%(17/1348)	0.008	0.041
451N	B	13.0%(49/378)	14.7%(198/1348)	0.962	1
451S	C	42.0%(50/119)	38.5%(663/1721)	0.992	1
451S	G	14.3%(2/14)	6.1%(2/33)	0.395	1
451T	B	2.1%(8/378)	0.0%(0/1348)	<0.0001	<0.0001
452G	B	0.8%(3/384)	0.1%(2/1374)	0.074	0.542
452G	G	7.7%(1/13)	0.0%(0/33)	0.298	1
452K	B	1.0%(4/384)	0.9%(13/1374)	0.535	1
452S	B	3.4%(13/384)	0.3%(4/1374)	<0.0001	<0.0001
453I	B	0.8%(3/384)	0.2%(3/1399)	0.12	0.625
453I	G	15.4%(2/13)	0.0%(0/33)	0.093	1
453Ins	C	25.7%(35/136)	24.0%(413/1722)	0.3561	1
453L	B	18.5%(71/384)	7.1%(99/1399)	<0.0001	<0.0001
453L	C	3.4%(4/119)	10.4%(179/1722)	0.999	1
453T	B	4.4%(17/384)	4.7%(66/1399)	0.699	1
453T	C	21.8%(26/119)	3.1%(53/1722)	<0.0001	<0.0001
453T	01_AE	40.9%(9/22)	14.1%(157/1112)	0.026	0.386
469I	B	3.1%(12/383)	2.9%(46/1570)	0.508	1
469I	C	1.7%(2/116)	1.1%(19/1660)	0.411	1
469T	G	14.3%(2/14)	0.0%(0/33)	0.102	1
472P	01_AE	9.5%(2/21)	13.1%(121/927)	0.821	1
472S	B	0.9%(4/426)	1.6%(28/1795)	0.893	1
472S	C	1.8%(2/114)	7.6%(119/1560)	0.999	1
474L	B	0.5%(2/430)	0.3%(6/1811)	0.475	1
474P	B	0.7%(3/430)	1.7%(30/1811)	0.968	1
474P	01_AE	4.5%(1/22)	21.2%(218/1028)	0.996	1
486F	B	0.2%(1/433)	0.5%(9/1806)	0.884	1
486F	C	0.8%(1/119)	0.1%(1/1728)	0.126	1

*: We only examined the sequence datasets of subtypes B, C, G and CRF01_AE, which contained more than 10 sequences estimated to be partially or fully PI-resistant (see **Table 4.5**).

#: One-tailed Fisher's exact tests were performed on amino acid variants that occurred more than once in individual subtype datasets. For each HIV-1 subtype, the obtained p-values were adjusted using multiple testing corrections via the false discovery rate

approach (software: Matlab 2013a, also see reference [44]). Amino acid variants are colored dark gray if their p-values and adjusted p-values were less than 0.05. Amino acid variants are colored gray if their p-values were less than 0.05 but adjusted p-values were above 0.05. Amino acid variants whose p-values are equal to 1 are not shown.

4.6 References

1. Larrouy L, Vivot A, Charpentier C, Benard A, Visseaux B, Damond F, *et al.* Impact of gag genetic determinants on virological outcome to boosted lopinavir-containing regimen in HIV-2-infected patients. *AIDS* 2013,**27**:69-80.
2. Larrouy L, Chazallon C, Landman R, Capitant C, Peytavin G, Collin G, *et al.* Gag mutations can impact virological response to dual-boosted protease inhibitor combinations in antiretroviral-naive HIV-infected patients. *Antimicrob Agents Chemother* 2010,**54**:2910-2919.
3. Wensing AM, van Maarseveen NM, Nijhuis M. Fifteen years of HIV Protease Inhibitors: raising the barrier to resistance. *Antiviral Res* 2010,**85**:59-74.
4. Fun A, Wensing AM, Verheyen J, Nijhuis M. Human Immunodeficiency Virus Gag and protease: partners in resistance. *Retrovirology* 2012,**9**:63.
5. Larrouy L, Charpentier C, Landman R, Capitant C, Chazallon C, Yeni P, *et al.* Dynamics of gag-pol minority viral populations in naive HIV-1-infected patients failing protease inhibitor regimen. *AIDS* 2011,**25**:2143-2148.
6. Gupta RK, Kohli A, McCormick AL, Towers GJ, Pillay D, Parry CM. Full-length HIV-1 Gag determines protease inhibitor susceptibility within in vitro assays. *AIDS* 2010,**24**:1651-1655.
7. Knops E, Kemper I, Schulter E, Pfister H, Kaiser R, Verheyen J. The evolution of protease mutation 76V is associated with protease mutation 46I and gag mutation 431V. *AIDS* 2010,**24**:779-781.
8. Aoki M, Venzon DJ, Koh Y, Aoki-Ogata H, Miyakawa T, Yoshimura K, *et al.* Non-cleavage site gag mutations in amprenavir-resistant human immunodeficiency virus type 1 (HIV-1) predispose HIV-1 to rapid acquisition of amprenavir resistance but delay development of resistance to other protease inhibitors. *J Virol* 2009,**83**:3059-3068.
9. Chang MW, Oliveira G, Yuan J, Okulicz JF, Levy S, Torbett BE. Rapid deep sequencing of patient-derived HIV with ion semiconductor technology. *J Virol Methods* 2013,**189**:232-234.
10. Parry CM, Kolli M, Myers RE, Cane PA, Schiffer C, Pillay D. Three residues in HIV-1 matrix contribute to protease inhibitor susceptibility and replication capacity. *Antimicrob Agents Chemother* 2011,**55**:1106-1113.
11. Gatanaga H, Suzuki Y, Tsang H, Yoshimura K, Kavlick MF, Nagashima K, *et al.* Amino acid substitutions in Gag protein at non-cleavage sites are indispensable for the development of a high multitude of HIV-1 resistance against protease inhibitors. *J Biol Chem* 2002,**277**:5952-5961.
12. Seclen E, Gonzalez Mdel M, Corral A, de Mendoza C, Soriano V, Poveda E. High prevalence of natural polymorphisms in Gag (CA-SP1) associated with reduced response to Bevirimat, an HIV-1 maturation inhibitor. *AIDS* 2010,**24**:467-469.
13. Malet I, Roquebert B, Dalban C, Wirden M, Amellal B, Agher R, *et al.* Association of Gag cleavage sites to protease mutations and to virological response in HIV-1 treated patients. *J Infect* 2007,**54**:367-374.
14. Ghosn J, Delaugerre C, Flandre P, Galimand J, Cohen-Codar I, Raffi F, *et al.* Polymorphism in Gag gene cleavage sites of HIV-1 non-B subtype and virological outcome of a first-line lopinavir/ritonavir single drug regimen. *PLoS One* 2011,**6**:e24798.
15. Mo H, Parkin N, Stewart KD, Lu L, Dekhtyar T, Kempf DJ, *et al.* Identification and structural characterization of I84C and I84A mutations that are associated with high-level resistance to human immunodeficiency virus protease inhibitors and impair viral replication. *Antimicrob Agents Chemother* 2007,**51**:732-735.
16. Nijhuis M, Wensing AM, Bierman WF, de Jong D, Kagan R, Fun A, *et al.* Failure of treatment with first-line lopinavir boosted with ritonavir can be explained by novel resistance pathways with protease mutation 76V. *J Infect Dis* 2009,**200**:698-709.

17. van Maarseveen NM, Andersson D, Lepsik M, Fun A, Schipper PJ, de Jong D, *et al.* Modulation of HIV-1 Gag NC/p1 cleavage efficiency affects protease inhibitor resistance and viral replicative capacity. *Retrovirology* 2012,**9**:29.
18. Kolli M, Stawiski E, Chappey C, Schiffer CA. Human immunodeficiency virus type 1 protease-correlated cleavage site mutations enhance inhibitor resistance. *J Virol* 2009,**83**:11027-11042.
19. Larrouy L, Lambert-Niclot S, Charpentier C, Fourati S, Visseaux B, Soulie C, *et al.* Positive impact of HIV-1 gag cleavage site mutations on the virological response to darunavir boosted with ritonavir. *Antimicrob Agents Chemother* 2011,**55**:1754-1757.
20. Banke S, Lillemark MR, Gerstoft J, Obel N, Jorgensen LB. Positive selection pressure introduces secondary mutations at Gag cleavage sites in human immunodeficiency virus type 1 harboring major protease resistance mutations. *J Virol* 2009,**83**:8916-8924.
21. Bally F, Martinez R, Peters S, Sudre P, Telenti A. Polymorphism of HIV type 1 gag p7/p1 and p1/p6 cleavage sites: clinical significance and implications for resistance to protease inhibitors. *AIDS Res Hum Retroviruses* 2000,**16**:1209-1213.
22. Cote HC, Brumme ZL, Harrigan PR. Human immunodeficiency virus type 1 protease cleavage site mutations associated with protease inhibitor cross-resistance selected by indinavir, ritonavir, and/or saquinavir. *J Virol* 2001,**75**:589-594.
23. Verheyen J, Litau E, Sing T, Daumer M, Balduin M, Oette M, *et al.* Compensatory mutations at the HIV cleavage sites p7/p1 and p1/p6-gag in therapy-naive and therapy-experienced patients. *Antivir Ther* 2006,**11**:879-887.
24. Knops E, Brakier-Gingras L, Schulter E, Pfister H, Kaiser R, Verheyen J. Mutational patterns in the frameshift-regulating site of HIV-1 selected by protease inhibitors. *Med Microbiol Immunol* 2012,**201**:213-218.
25. Nijhuis M, van Maarseveen NM, Lastere S, Schipper P, Coakley E, Glass B, *et al.* A novel substrate-based HIV-1 protease inhibitor drug resistance mechanism. *PLoS Med* 2007,**4**:e36.
26. Lambert-Niclot S, Flandre P, Malet I, Canestri A, Soulie C, Tubiana R, *et al.* Impact of gag mutations on selection of darunavir resistance mutations in HIV-1 protease. *J Antimicrob Chemother* 2008,**62**:905-908.
27. Prado JG, Wrin T, Beauchaine J, Ruiz L, Petropoulos CJ, Frost SD, *et al.* Amprenavir-resistant HIV-1 exhibits lopinavir cross-resistance and reduced replication capacity. *AIDS* 2002,**16**:1009-1017.
28. Brann TW, Dewar RL, Jiang MK, Shah A, Nagashima K, Metcalf JA, *et al.* Functional correlation between a novel amino acid insertion at codon 19 in the protease of human immunodeficiency virus type 1 and polymorphism in the p1/p6 Gag cleavage site in drug resistance and replication fitness. *J Virol* 2006,**80**:6136-6145.
29. Maguire MF, Guinea R, Griffin P, Macmanus S, Elston RC, Wolfram J, *et al.* Changes in human immunodeficiency virus type 1 Gag at positions L449 and P453 are linked to I50V protease mutants in vivo and cause reduction of sensitivity to amprenavir and improved viral fitness in vitro. *J Virol* 2002,**76**:7398-7406.
30. Myint L, Matsuda M, Matsuda Z, Yokomaku Y, Chiba T, Okano A, *et al.* Gag non-cleavage site mutations contribute to full recovery of viral fitness in protease inhibitor-resistant human immunodeficiency virus type 1. *Antimicrob Agents Chemother* 2004,**48**:444-452.
31. Roquebert B, Malet I, Wirden M, Tubiana R, Valantin MA, Simon A, *et al.* Role of HIV-1 minority populations on resistance mutational pattern evolution and susceptibility to protease inhibitors. *AIDS* 2006,**20**:287-289.
32. Kolli M, Lastere S, Schiffer CA. Co-evolution of nelfinavir-resistant HIV-1 protease and the p1-p6 substrate. *Virology* 2006,**347**:405-409.
33. Kaufmann GR, Suzuki K, Cunningham P, Mukaide M, Kondo M, Imai M, *et al.* Impact of HIV type 1 protease, reverse transcriptase, cleavage site, and p6 mutations on the virological response to quadruple therapy with saquinavir, ritonavir, and two nucleoside analogs. *AIDS Res Hum Retroviruses* 2001,**17**:487-497.
34. Yates PJ, Hazen R, St Clair M, Boone L, Tisdale M, Elston RC. In vitro development of resistance to human immunodeficiency virus protease inhibitor GW640385. *Antimicrob Agents Chemother* 2006,**50**:1092-1095.
35. Lastere S, Dalban C, Collin G, Descamps D, Girard PM, Clavel F, *et al.* Impact of insertions in the HIV-1 p6 PTAPP region on the virological response to amprenavir. *Antivir Ther* 2004,**9**:221-227.

36. Knops E, Daumer M, Awerkiew S, Kartashev V, Schulter E, Kutsev S, *et al.* Evolution of protease inhibitor resistance in the gag and pol genes of HIV subtype G isolates. *J Antimicrob Chemother* 2010,**65**:1472-1476.
37. Rossi AH, Rocco CA, Mangano A, Sen L, Aulicino PC. Sequence variability in p6 gag protein and gag/pol coevolution in human immunodeficiency type 1 subtype F genomes. *AIDS Res Hum Retroviruses* 2013,**29**:1056-1060.
38. Libin P, Beheydt G, Deforche K, Imbrechts S, Ferreira F, Van Laethem K, *et al.* RegaDB: community-driven data management and analysis for infectious diseases. *Bioinformatics* 2013,**29**:1477-1480.
39. Van Laethem K, Schrooten Y, Dedecker S, Van Heeswijck L, Deforche K, Van Wijngaerden E, *et al.* A genotypic assay for the amplification and sequencing of gag and protease from diverse human immunodeficiency virus type 1 group M subtypes. *J Virol Methods* 2006,**132**:181-186.
40. Maes B, Schrooten Y, Snoeck J, Derdelinckx I, Van Ranst M, Vandamme AM, *et al.* Performance of ViroSeq HIV-1 Genotyping System in routine practice at a Belgian clinical laboratory. *J Virol Methods* 2004,**119**:45-49.
41. Liu TF, Shafer RW. Web resources for HIV type 1 genotypic-resistance test interpretation. *Clinical infectious diseases* 2006,**42**:1608-1618.
42. Van Laethem K, De Luca A, Antinori A, Cingolani A, Perna CF, Vandamme AM. A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients. *Antivir Ther* 2002,**7**:123-129.
43. Li G, Verheyen J, Rhee SY, Voet A, Vandamme AM, Theys K. Functional conservation of HIV-1 gag: implications for rational drug design. *Retrovirology* 2013,**10**:126.
44. Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002,**64**:479-498.
45. de Oliveira T, Engelbrecht S, Janse van Rensburg E, Gordon M, Bishop K, zur Megede J, *et al.* Variability at human immunodeficiency virus type 1 subtype C protease cleavage sites: an indication of viral fitness? *J Virol* 2003,**77**:9422-9430.
46. Martins AN, Arruda MB, Pires AF, Tanuri A, Brindeiro RM. Accumulation of P(T/S)AP late domain duplications in HIV type 1 subtypes B, C, and F derived from individuals failing ARV therapy and ARV drug-naive patients. *AIDS Res Hum Retroviruses* 2011,**27**:687-692.

Chapter 5

A new ensemble coevolution system for detecting HIV-1 protein coevolution

"Never tell yourself that something it's too hard."

— Mitch Albom

This chapter is adapted from my article:

Guangdi Li, Kristof Theys, Jens Verheyen, Andrea-Clemencia Pineda-Peña, Ricardo Khouri, Supinya Piampongsant, Mónica Eusébio, Jan Ramon, Anne-Mieke Vandamme. A new ensemble coevolution system for detecting HIV-1 protein coevolution.

I proposed the idea, designed the software and drafted the manuscript. The improvement of the paper was supported with substantial help from Prof. Anne-Mieke Vandamme, Kristof Theys and Prof. Jan Ramon, as well as advices and corrections from other coauthors. I sincerely thank Fossie Ferreira and Jasper Edgar Neggers for technical assistance and valuable contributions to the analysis.

5.1 Summary

A key challenge in the field of HIV-1 protein evolution is the identification of coevolving amino acids at the molecular level. In the past decades, many sequence-based methods have been designed to detect position-specific coevolution within and between different proteins. However, an ensemble coevolution system that integrates different methods to improve the detection of HIV-1 protein coevolution has not been developed. We integrated 27 sequence-based prediction methods published between 2004 and 2013 into an ensemble coevolution system. This system allowed combinations of different sequence-based methods for coevolution predictions. Using HIV-1 protein structures and experimental data, we evaluated the performance of individual and combined sequence-based methods in the prediction of HIV-1 intra- and inter-protein coevolution. We showed that sequence-based methods clustered according to their methodology, and a combination of four methods outperformed any of the 27 individual methods. This four-method combination estimated that HIV-1 intra-protein coevolving positions were mainly located in functional domains and physically contacted with each other in the protein tertiary structures. In the analysis of HIV-1 inter-protein coevolving positions between Gag and protease, protease drug resistance positions near the active site mostly coevolved with Gag cleavage positions (V128, S373-T375, A431, F448-P453) and Gag C-terminal positions (S489-Q500) under selective pressure of protease inhibitors. This study presents a new ensemble coevolution system which detects position-specific coevolution using combinations of 27 different sequence-based methods. Our findings highlight key coevolving residues within HIV-1 structural proteins and between Gag and protease, shedding light on HIV-1 intra- and inter-protein coevolution.

5.2 Introduction

Recent structural analysis showed that the fullerene core of HIV-1 particles is formed by capsid hexamers and pentamers through both intra- and inter-protein interactions [1]. HIV-1 capsid protein is encoded by the *gag* gene, which contains matrix, capsid, p2, nucleocapsid, p1 and p6 domains. In a spherical shell of an immature virus, Gag polyproteins are arranged radially in a curved hexameric lattice bound together by protein interactions [2]. The HIV-1 matrix and capsid proteins are cleaved from Gag

and reorganized into tubular lattices of mature particles during the protease-mediated proteolytic processing [3]. Mutations near Gag cleavage sites (GCS) can affect the protease binding affinity [4], suggesting that HIV-1 intra- and inter-protein interactions play a key role during the viral life cycle. Previous sequence analyses have reported the association between human HLA alleles and Gag codons [5], intra-protein coevolution in capsid [6] and immunologically vulnerable sectors in Gag [7]. However, a systematic study of HIV-1 intra- and inter-protein coevolution of Gag and protease proteins is largely lacking.

Many studies have revealed position-specific coevolution in HIV-1 proteins using sequence-based methods [5, 6, 8-12]. For instance, coevolving positions were found to be proximal in capsid structure [6]. HIV-1 drug-resistance mutations in protease, reverse transcriptase and integrase tend to coevolve under the drug selective pressure [8-10, 13]. Important coevolving residues were also found in HIV-1 Env [11], Vif [12] and Gag [5]. To model coevolution within and between proteins [11, 14, 15], position-specific sequence analysis has been used to detect pairs of correlated amino acid positions, so-called statistical couplings [16] (also called co-variations [17] or correlated substitutions [18]). A deep understanding of genetically coevolving residues has enriched our insights in protein folding [17], protein-protein interaction [19], allosteric communication [20] and ligand binding [21] (see review [22]). Since the first sequence-based method was proposed in 1970 [23], more than 30 methods were published and most of them were based on the principle of information theory, physicochemical properties, molecular phylogenetics and Bayesian statistics [15, 22, 24]. Thanks to the increase of crystalized structures in public databases, the performance of sequence-based methods is usually evaluated based on structural information, such as protein contact map [25], because spatially proximate positions tend to coevolve [26] and sequence evolution is associated with structural dynamics [27]. Nevertheless, state-of-the-art methods in different studies showed significant variability, while evaluation of long-range coevolving residues continues to be difficult in most scenarios [15, 22, 24].

The supervised ensemble approach in statistics and machine learning aims at creating a robust method through the integration of multiple predictive models [28]. It relies on the philosophy that the aggregation of information from several sources is usually

superior to a single individual source for decision-making (e.g. jury, peer-review, voting for political candidates) [28]. Well-known ensemble methods such as random forest [29] and AdaBoost [30] provide robust predictions with outstanding performance in many applications. Other ensemble methods have also been designed for solving various problems [31-33]. For instance, the ensemble machine system XCS was made to improve self-adaptation of evolutionary algorithms [31]. While more than 27 sequence-based methods have been proposed for position-specific coevolution prediction, an ensemble coevolution system that integrates multiple methods to improve the prediction of HIV protein coevolution has not been investigated.

Here, we present the first ensemble coevolution system (ECS) to detect HIV-1 position-specific coevolution by integrating 27 sequence-based methods published between 2004 and 2013 (**Table 5.1**). This new software platform allows for parallel coevolution predictions and systematic combinations of sequence-based methods. We collected extensive HIV-1 sequences and experimental and clinical data to evaluate the performance of individual methods and combinations of methods. Using our coevolution system, we identified combinatorial approaches with superior performance at predicting HIV-1 coevolution. We thereafter investigated intra- and inter-protein coevolving positions in HIV-1 Gag and protease using an optimized combinatorial approach that integrated four sequence-based methods.

Table 5.1: Summary of 27 sequence-based methods in our ensemble coevolution system.

Methods*	Statisitcal methodology	Updated	Ref
ASC/APC	Mutual information	2007	[34]
BN	Bayesian network	2007	[35]
CTMP	Continuous-time Markov model, phylogenetic tree	2007	[36]
CoMap	Compensation coefficient, phylogenetic tree	2007	[37]
Complementary	AA complementary matrix, Pearson coefficient	2006	[38]
CMPro	2D recursive neural networks	2012	[39]
DCA	Maximum entropy model	2011	[25, 26]
DNcon	Deep network, Bolzmann machines	2012	[40]
GREMLIN	Maximum entropy model	2013	[41]
Interdependency	Entropy, mutual information	2004	[42]
LogR	Bayesian networks, APC	2010	[43]
MI	Mutual information	2012	[44-46]
MIBP	Mutual information,physicochemical properties	2011	[47]

Mutagenetic	Maximum likelihood mixed trees	2005	[10]
NBZPX2	Normal binary, ZRES	2012	[46]
NCPS	Mutual information, sequence similarity	2009	[48]
NNcon	Neural networks	2009	[49]
PCC	Mutual information, Pearson's coefficients	2010	[18]
PSICOV	Sparse inverse covariance	2012	[50]
PhysicoMI	Mutual information, physicochemical properties	2012	[6]
PhyCMAP	Random forest, integer linear programming	2013	[51]
RCW	Mutual information	2007	[52]
Spidermonkey	MCMC Bayesian network, phylogenetic tree	2008	[53]
SCA	Statistical free energy couplings	2009	[54]
SVMcon	Support vector machine	2006	[55]
ZRES	Mutual information	2009	[56]

*: A comprehensive description of the methodology and our experimental settings are provided in section 2 of **Text S1**.

5.3 Materials and Methods

HIV-1 protein sequence datasets for sequence-based coevolution prediction

As of February 2013, we retrieved 3171 HIV-1 subtype B *gag* and protease nucleotide sequences from the Los Alamos HIV database (http://www.hiv.lanl.gov) (HXB2 nucleotide positions: 1186-2549, one sequence per patient). For each Gag and protease protein, we aligned sequences against the HXB2 reference and manually curated the alignment using Seaview V4.3 [57]. To improve sequence quality, we used the criteria described in our recent study [58] to remove duplicates and sequences with any hypermutation, stop codon, ambiguous nucleotide or subtype misclassification. Afterwards, patient treatment information of the retrieved sequences was obtained from the corresponding sequence publications. Sequence data obtained from treatment-naive patients were used to detect intra-protein statistical couplings given that wild-type HIV-1 protein structures were used for evaluation. Sequence data obtained from patients receiving protease inhibitor (PI) treatment were used to detect inter-protein statistical couplings given that HIV-1 clinical datasets with PI treatment information were used for evaluation. Overall, we obtained five intra-protein sequence datasets: matrix (n=605), capsid (n=656), nucleocapsid (n=768), p6 (n=1030), protease (n=1762), as well as two inter-protein sequence datasets, protease-p6 (n=788) and protease-GCS (Gag cleavage sites) (n=292).

Sequence-based statistical methods for position-specific coevolution predictions

We integrated 27 known sequence-based statistical methods (**Text S1**) into one software platform for position-specific coevolution predictions. Summarized in **Table 5.1**, these methods were mainly designed based on the principles of information theory, phylogenetic analysis, parametric or non-parametric statistical tests, Bayesian maximum likelihood and codon substitution models. Given the inputs of multiple sequence alignments (MSAs) and phylogenetic trees, sequence-based methods predict coevolving residues and rank them according to the method-specific measurements with either parametric or non-parametric statistical tests (**Text S1**). The predictions were ranked according to each method. Parameter settings used in our study were either default or optimized according to method manuals or publications (**Text S1**). To prepare the inputs of the phylogenetic-based methods, we constructed unrooted maximum likelihood phylogenetic trees using the following procedure. Given the nucleotide MSAs, neighbor-joining phylogenetic trees were obtained by IQPNNI V 3.3 [59] (nucleotide substitution model: general time reversible (GTR) model, bootstrap resampling: 1000 replicates). These neighbor-joining phylogenetic trees were used as starting trees in RAxML V7.0.4, which subsequently optimized the unrooted maximum likelihood phylogenetic trees (nucleotide substitution model: GTRGAMMA, 100 bootstrap replicates) [60].

HIV-1 protein structural and experimental datasets for evaluating predictive performance of sequence-based methods

We retrieved PDB data of HIV-1 proteins from the RCSB Protein Data Bank (www.pdb.org). The quality of crystalized structures was assessed using PDBREPORT [61] (default parameters). The PDB dataset included: 1HIW (matrix), 3H4E (capsid), 1A1T (nucleocapsid), 2C55 (p6) and 1TW7 (protease). We also collected extensive experimental and clinical data of PI-associated Gag-protease mutations from literature, which was queried in PubMed using the keywords “HIV Gag mutation”, “HIV Gag protease”, “HIV protease mutations Gag”, “HIV Gag evolution” or “HIV protease cleavage”. References in primary studies and reviews were also searched. The data is summarized in **Table S 5.1**.

True positives for intra-protein coevolving positions were assessed according to their proximity in protein contact maps. To construct contact maps for each protein, Euclidean distances between the C_β atoms of residue pairs were calculated given the atomic coordinates in PDB [51]. In cases where a HIV-1 protein has multiple functional domains (e.g. matrix, capsid, protease), Euclidean distances between residue pairs were calculated within and between functional domains and the minimum value for each pair was used for assignment [25]. The predicted intra-protein couplings were assigned as true positives if they were long-range pairs of residues in contact: (1) at least 6 amino acids apart in the sequence [51]; (2) not located at the same alpha-helix or beta-strand secondary structures [49] and (3) less than 8 Å between residue pairs on the protein contact map [25]. The predicted intra-protein couplings, which had residues less than 6 amino acids apart in the sequence or were located in the same alpha-helix or beta-strand secondary structures, were not counted during the evaluation. Above criteria were set to evaluate long-range coevolving positions in protein tertiary structures by not counting predictions of neighboring AA positions.

For the protease-p6 and protease-GCS coevolution, the predicted inter-protein residue pairs were considered as true positives if any corresponding Gag-protease mutation patterns were reported in the experimental and clinical datasets (**Table S 5.1**). For each row of multiple residue patterns in **Table S 5.1**, pairwise combinations of protease-p6 or protease-GCS residues were used for the validation of true positives.

For both intra- and inter-protein predictions, false positives were the couplings in the top-ranked long-range predictions that were not identified as true positives. We did not evaluate negative predictions because the sequence-based methods were not designed to predict residue positions that are not coevolving [22].

Statistical measurements for method evaluation

Predictions of sequence-based methods were evaluated by five statistical measurements.

Precision-recall curve (AUC): For intra- and inter-protein coevolution predictions, we assessed the area under the precision-recall curve (AUC) [62] as the relative effectiveness of sequence-based methods. Optimized by the binomial model, an

unbiased estimator of AUC was calculated by taking into account biases introduced by small sample sizes and class imbalance in favor of negative examples [62]. Notably, AUC is independent of the cutoffs of the top-ranked long-range couplings and is equal to one if all the true positives are ranked higher than the false positives.

Accuracy: For intra- and inter-protein coevolution predictions, accuracy was calculated as the number of true positives divided by the total number of top-ranked predictions [40, 55, 63]. Particularly, the accuracy of the L/2 or L top-ranked predictions was evaluated, where L was the number of residue positions in the MSA input. In most instances, the cutoff for positive predictions of coevolving pairs of residues or couplings was set to the L top-ranked couplings. In some instances (mentioned specifically), it was set to the L/2 top-ranked predictions [63]. Thus, positive predictions for coevolution are the L top-ranked couplings, unless it is specified that L/2 is used as a cutoff.

Harmonic distance: For intra-protein coevolution predictions, the harmonic distance X_d was measured as a weighted harmonic average difference between the Euclidean distance distribution of the predicted couplings and the all-pair Euclidean distances [51, 63]. Being popular in Critical Assessment of Protein Structure Prediction (CASP), the harmonic distance X_d is defined as: $X_d = \sum_{n=1}^{15} (P(d_n) - P(a_n))/n$, where $P(d_n)$ is the percentage of predicted couplings with Euclidean distances between 4(n−1) and 4n, $P(a_n)$ is the percentage of all contact pairs with Euclidean distances between 4(n−1) and 4n [51]. A higher value of the harmonic distance X_d indicates a better prediction performance of a method.

Average Euclidean distance: For intra-protein coevolution predictions, average Euclidean distance was measured for the top-ranked long-range couplings using the Cβ-Cβ Euclidean distances [25]. It is defined as: $\sum_{i=1}^{L} Dist(C_i, C_i)/L$, where L is the number of top-ranked couplings, C_i and C_i are two residue positions in the i^{th} top-ranked long-range coupling. For evaluation purposes, the number of top-ranked couplings predicted by individual methods was set to L/2 or L [63]. A lower value of average Euclidean distance indicates better prediction performance of a method.

Jaccard and association coefficients: To quantify the predictive heterogeneity of sequence-based methods, Jaccard and association coefficients were calculated between the top-ranked long-range couplings predicted by different sequence-based methods. Given two coupling sets X and Y, Jaccard and association coefficients are defined as $|X \cap Y| / |X \cup Y|$ and $|X \cap Y| / \min(|X|, |Y|)$, respectively [64].

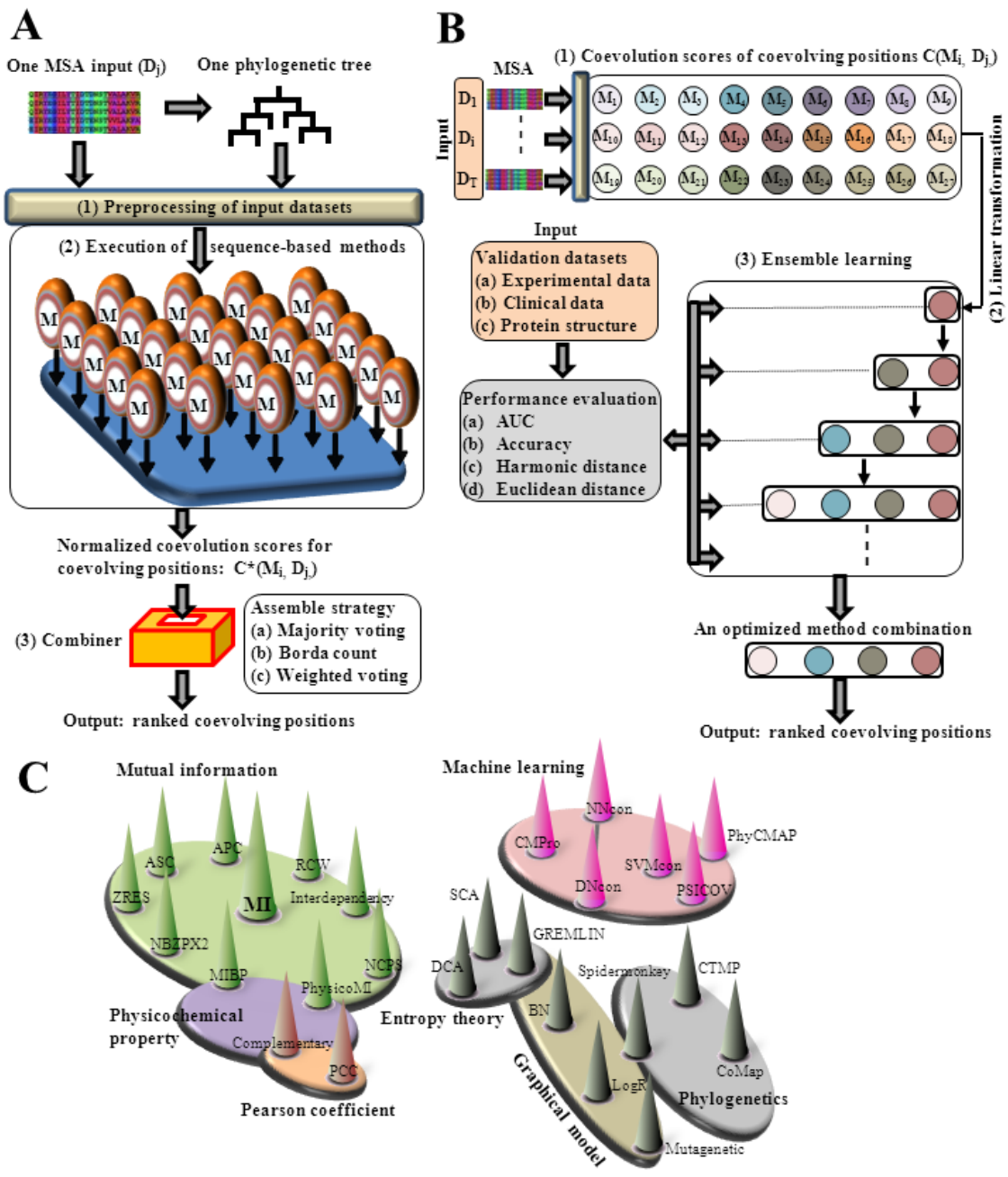


Figure 5.1: Schematic view of ensemble coevolution system
(A) Workflow of coevolution prediction. Input data: a multiple sequence alignment dataset D_j and one phylogenetic tree constructed using D_j. (1) Preprocessing of input

datasets, the method-specific input formats are preprocessed and imported into individual sequence-based methods $M_i (i = 1, ..., 27)$. (2) Execution of sequence-based methods, sequence-based methods are applied to predict coevolving positions using the parallel computation. Each method predicts a list of coevolving positions with estimated coevolution scores. Given the sequence-based method M_i and the sequence dataset D_j, coevolution scores of coevolving positions are normalized and exported into the matrix $C^*(M_i, D_j)$. (3) Combiner, given a chosen combination of sequence-based methods, coevolution scores of predicted coevolving positions are assembled through the combiner, which provides the assemble strategies such as majority voting, Borda count and weighted voting. Coevolving positions are ranked and exported as outputs with corresponding coevolution scores.
(B) Workflow of our procedures that optimize the combination of sequence-based methods. Input data: inputs of multiple MSAs are processed by sequence-based methods (see (A)). The validation datasets (e.g. experimental and clinical data) are also prepared for the method evaluation. Coevolution scores of ranked coevolving pairs in $C(M_i, D_j)$ are collected after applying the sequence-based method M_i to the sequence dataset D_j. (1) Linear transformation, the coevolution scores are linearly transformed between 0 and 1. (2) Ensemble learning, a heuristic algorithm identifies the combination of sequence-based methods with improved prediction performance (Text S1). Each circle represents a single method and the combination of different methods is demonstrated in a group of colored circles. Using the validation datasets, prediction performance is evaluated (e.g. AUC) for the ranked statistical couplings assembled from the corresponding method combination. When adding a new method will not improve the prediction performance, the learning procedure stops and an optimized method combination is identified. Using the identified method combination, coevolving pairs are predicted as in (A) and returned as outputs.
(C) Correlation-based networks of sequence-based methods. Seven major methodologies are summarized, including mutual information, machine learning (random forest, support vector machine, neural networks), Pearson coefficient, entropy theory, graphical models (Bayesian networks, singly connected spanning trees, mutagenetic trees), phylogenetic models and physicochemical property models. Methods are represented by cones (see abbreviations in **Table 5.1**) and the same color is given to sequence-based methods designed from similar methodologies (e.g. APC used MI as a part of its design, phylogenetic trees are used in Spidermonkey, CTMP and CoMap).

Ensemble coevolution system (ECS)

To provide robust position-specific coevolution predictions, we designed an ensemble coevolution system by integrating 27 sequence-based methods published in the last decade (**Table 5.1**). Inspired by the ensemble principle [65], ECS's workflow includes: (1) inputs of MSAs and their corresponding phylogenetic trees, (2) execution of sequence-based methods, (3) a method combiner which integrates prediction results from different methods. **Figure 5.1**A shows the schematic overview of ECS and its model is described as follows. Suppose we have a set of sequence-

based methods, denoted as $M = \{M_1, M_2, ..., M_N\}$ and multiple sequence datasets, denoted as: $D = \{D_1, ..., D_T\}$, where N is the number of methods (N = 27 in our study) and T is the number of sequence datasets (T = 7 in our study).

Execution of sequence-based methods: Given a dataset D_j, the method M_i quantifies a coevolution score for the statistical coupling between the nth and the mth positions ($n, m \in \{1, ..., L\}$), where L is the number of amino acid positions in D_j. The higher the score, the higher the statistical significance based on the method-specific measurements. This process generates a scoring matrix $C(M_i, D_j)$ which has at most L×L pairs. The coevolution scores in $C(M_i, D_j)$ are then linearly transformed between 0 and 1 ($C^*(M_i, D_j) = [C(M_i, D_j) - \min(C)] / [\max(C) - \min(C)]$), where the higher the score, the higher the statistical significance. For each MSA evaluated by each method, the normalized coevolution scores in the scoring matrix are ranked with the highest score being the top ranked (see section 2 in **Text S1**).

Method combiner: Users can choose any individual methods to combine, or use three implemented assemble strategies (majority voting, Borda count, weighted voting)[65]. For the majority voting, the combiner outputs the predicted coevolving residues if they were predicted in the (L or L/2) top-ranked predictions by more than half of the 27 sequence-based methods. For the Borda count, the combiner outputs only the coevolving residues if they were predicted in the (L or L/2) top-ranked coupling predictions by all the 27 sequence-based methods. For weighted voting, ranking is done after collecting the weighted votes. The weighted votes are collected as follows:

Suppose a combination of methods is denoted by Ω, $|\Omega|$ is the number of methods in the method combination Ω, and w_i is the weight of sequence-based method M_i contributed to the coevolution scores. All methods contribute equally when every w_i equals to 1. The normalized coevolution scores $C^*_{n,m}(\Omega, D_j)$ is defined as:

$$C^*_{n,m}(\Omega, D_j) = \frac{1}{|\Omega|} \sum_{M_i \in \Omega} w_i \times C^*_{n,m}(M_i, D_j)$$

$C^*_{n,m}(\Omega, D_j)$ is thereafter ranked and exported as outputs. Notably, Ω can either contain a single method or a combination of methods, which can be selected based on the performance evaluation (see next section).

Identification of method combinations using a heuristic algorithm

Using validation datasets to evaluate the method performance, we proposed a heuristic algorithm to optimize a method combination. Given a performance measurement f (e.g. AUC), $f(C^*(\Omega, D_j))$ measures the statistical performance of the method combination Ω applied to the dataset D_j. To identify an optimized combination of methods, an objective function $F(\Omega, D)$ is defined by a linear function [66]:

$$F(\Omega, D) = \sum_{j=1}^{T} u_j \times f(C^*_{n,m}(\Omega, D_j))$$

Where u_j is the weight of the training dataset D_j contributed to the objective function. All datasets are treated equally if every u_j equals to 1.

Based on the objective function, an optimized combination of methods, denoted as Ω^+, is obtained by $\Omega^+ = \max_{\Omega \subseteq M} F(\Omega, D)$. Given the 27 known sequence-based methods, we aimed at identifying a method combination Ω^+ to achieve a high prediction performance, preferably combining only a small number of methods. The reason for this is twofold. Firstly, some coevolution methods are computationally heavy. Secondly, it is hard to implement and apply an ensemble system integrating many complex methods. To simplify the optimization procedure, we also assumed that all training datasets contributed equally ($u_j = 1$) and sequence-based methods contribute equally in a method combiner when selected (w_i equals to 1 or 0). Inspired by the forward selection and backward elimination approach [67], we designed a heuristic algorithm to identify the smallest method subset that maximizes the objective function. Text S1 clarifies this heuristic algorithm with more mathematical details. Here we provide an overview of the underlying principle.

Our heuristic algorithm begins with the independent predictions of the 27 sequence-based methods applied on the MSA inputs (**Figure 5.1B**). For each method with a MSA input, statistical couplings in the scoring matrix are ranked according to the method-specific significance measurements (**Text S1**, Section 2). In the next step, the forward selection each time visits all methods but only adds the method with the largest increase in performance into the method subsets and assembles the coupling predictions for evaluation. The procedure ends when adding a method does not further improve the best performance score. Similar to forward selection, the backward elimination is performed (see **Text S1**). To evaluate the performance of the score, AUC is used because it is a statistical measurement independent of the cutoffs of the top-ranked predictions.

5.4 Results

Estimate HIV-1 coevolution using a new ensemble coevolution system (ECS)

From the Los Alamos database, we retrieved 3171 nucleotide sequences of HIV-1 subtype B Gag and protease, resulting in five intra-protein datasets (matrix, capsid, nucleocapsid, p6, protease) and two inter-protein datasets (protease-p6, protease-GCS). We calculated protein contact maps based on the Euclidian distance between amino acids in the protein structures of matrix, capsid, nucleocapsid, p6 and protease. A Euclidian distance of less than 8 Å between residue pairs was considered as a biological measure of intra-protein coevolution [25]. We also performed a literature search of associated Gag and protease residues to identify inter-protein couplings confirmed by experimental and clinical studies. These data obtained from protein structure and literature review was used to validate true positive predictions of statistical couplings generated by sequence-based methods. We then designed an ensemble coevolution system (ECS) which integrates 27 sequence-based methods published between 2004 and 2013 (**Figure 5.1**, **Table 5.1**). Thereafter, we designed a heuristic algorithm to optimize the combination of sequence-based methods, which were evaluated by AUC (see Methods). Given our seven HIV-1 sequence datasets, this heuristic algorithm identified an optimized method combination, so-called CNPR, for the prediction of HIV-1 intra- and inter-protein coevolution (see section 1 of Text

S1). This CNPR combination comprised of four known methods (CMPro [39], NCPS [48], PhyCMAP [51] and RCW [52]), weighted equally (see section 1 in **Text S1**).

CNPR outperforms 27 known sequence-based methods in detecting HIV-1 coevolution

We found that CNPR outperformed each of the 27 sequence-based methods in the prediction of HIV-1 intra- and inter-protein coevolution using four statistical measurements (**Figure 5.2A**). All the 27 methods and the CNPR combination were evaluated and ranked for 7 HIV-1 sequence datasets, displayed in **Figure 5.2A**. Firstly, CNPR achieved the best average ranking (2.07) followed by CMPro (5.71) and PhyCMAP (6.87) based on the AUC measurement (**Table S 5.2**). Secondly, CNPR achieved the highest average accuracies for both the L/2 and L top-ranked predictions (average accuracy = 0.35, 0.27, respectively) (**Table S 5.2**). Comparing CNPR to the second best method NNcon, average accuracies for the L/2 and L top-ranked predictions increased by 0.061 (17.6%) and 0.031 (11.5%), respectively (**Table S 5.2, Table S 5.3**). Thirdly, we measured the harmonic distance X_d on the five intra-protein datasets. CNPR reached the second (X_d = 0.78) and the first ranking (X_d=0.66) on the L/2 and L top-ranked predictions, respectively (**Table S 5.2, Table S 5.4**). Fourthly, the L top-ranked long-range predictions of CNPR had the lowest average Euclidean distances (mean Euclidean distance: 11.52Å, 95% confidence interval: 4.64-20.85Å, **Figure 5.2B**). The L/2 top-ranked long-range predictions of CNPR had the second lowest average Euclidean distances (mean Euclidean distance: 10.14Å, 95% CI: 4.53-17.43Å).

Table 5.2: Performance of sequence-based methods in detecting HIV-1 protein coevolution

Method	Area-under-curve (AUC)							Accuracy		Harmonic distance		Euclidean distance	
	MA	CA	NC	p6	PR	p6-PR	CSM-PR	L/2	L	L/2	L	L/2	L
APC	0.57	0.55	0.59	0.71	0.57	0.62	0.66	0.108	0.086	0.039	0.027	17.38	18.6
ASC	0.56	0.53	0.59	0.75	0.59	0.63	0.62	0.15	0.117	0.051	0.028	16.41	18.69
BN	0.71	0.55	0.62	0.69	0.75	0.54	-*	0.059	0.052	0.009	0.008	19.94	20.13
CMPro	0.75	0.66	**0.85**	0.76	0.74	0.68	0.72	0.289	0.225	**0.166**	0.13	**10.05**	11.77
CTMP	0.54	0.52	-	-	0.57	0.69	-	0.033	0.033	0.004	0.004	16.98	16.98
CoMap	0.52	0.52	0.61	-	0.55	-	0.5	0.039	0.043	0.029	0.029	16.85	17.14
Complementary	0.52	0.52	0.57	0.54	0.53	0.53	0.55	0.04	0.047	0.008	0.003	19.08	20.01
DCA	0.55	0.55	0.59	0.78	0.51	0.64	0.67	0.092	0.071	0.03	0.023	17.43	18.45

DNcon	0.5	0.51	0.66	-	0.61	-	0.77	0.165	0.113	0.093	0.07	13.66	15.11
GREMLIN	0.56	0.54	0.6	0.81	0.6	0.6	0.63	0.138	0.095	0.04	0.024	17.14	18.77
Interdependency	0.63	0.58	0.68	-	0.66	-	-	0.073	0.07	0.028	0.026	18.4	18.58
LogR	0.55	0.54	0.54	0.8	0.55	0.58	0.55	0.114	0.083	0.024	0.015	18.44	19.32
MI	0.51	0.54	0.58	**0.84**	0.58	0.81	0.79	0.179	0.126	0.043	0.026	17.6	18.96
MIBP	0.57	0.5	0.57	0.67	0.53	0.62	0.7	0.045	0.053	0.021	0.023	17.8	18.12
Mutagenetic	0.53	0.66	0.71	-	0.64	0.86	0.6	0.159	0.159	0.027	0.027	19.13	19.13
NBZPX2	0.56	0.52	0.54	0.55	0.54	0.51	0.5	0.061	0.052	0.011	0.005	19.51	20.2
NCPS	0.58	0.51	0.54	0.83	0.56	0.86	0.83	0.17	0.116	0.018	0.011	19.37	20.27
NNcon	0.68	**0.72**	0.78	-	**0.78**	-	-	0.286	0.238	0.148	**0.132**	11.25	12.01
PCC	0.53	0.56	0.55	-	0.51	0.54	0.61	0.07	0.05	0.013	0	18.63	20.2
PSICOV	0.56	0.58	0.54	0.55	0.51	0.51	0.53	0.084	0.062	0.016	0.012	18.63	18.79
PhyCMAP	**0.76**	**0.7**	0.72	0.65	0.72	0.8	0.55	0.194	0.172	0.118	0.107	11.83	12.55
PhysicoMI	0.61	0.56	0.52	**0.84**	0.5	0.72	0.64	0.071	0.046	0.009	-0.001	20.46	21.06
RCW	0.54	0.53	0.58	0.82	0.56	0.8	0.78	0.123	0.109	0.044	0.032	16.88	18.12
SCA	0.54	0.54	0.56	0.53	0.58	0.77	0.77	0.157	0.108	0.027	0.016	18.26	19.08
SVMcon	0.71	**0.73**	0.67	-	**0.77**	-	-	0.246	0.183	0.14	0.111	11.42	12.65
Spidermonkey	0.58	0.55	0.63	0.67	0.52	0.51	0.57	0.065	0.057	0.018	0.01	18.89	19.77
ZRES	0.56	0.53	0.59	0.73	0.56	0.61	0.68	0.12	0.107	0.046	0.032	16.65	18.08
CNPR	0.75(2.5)	0.7(3.5)	0.83(2)	**0.84(1)**	0.77(2.5)	**0.87(1)**	**0.88(1)**	**0.347(1)**	**0.269(1)**	0.155(2)	**0.132(1.5)**	10.14(2)	**11.52(1)**

*: AUC was not evaluated due to the lack of long-range couplings predicted. For each column, the numbers in bold indicate methods with the best score among the 28 methods. The ranking of CNPR for each dataset is provided in brackets (see others in **Table S 5.2**). Ranking numbers in decimals are results from the average rankings (see examples in **Table S 5.2**). Four statiscal measurements (AUC, accuracy, harmonic distance, Euclidean distance) are defined in Methods. For the latter 3 methods, the L or L/2 top-ranked predictions were compared and the average scores over the 7 HIV-1 datasets were listed (see performance evaluation per method per dataset in **Table S 5.3**-**Table S 5.5**).

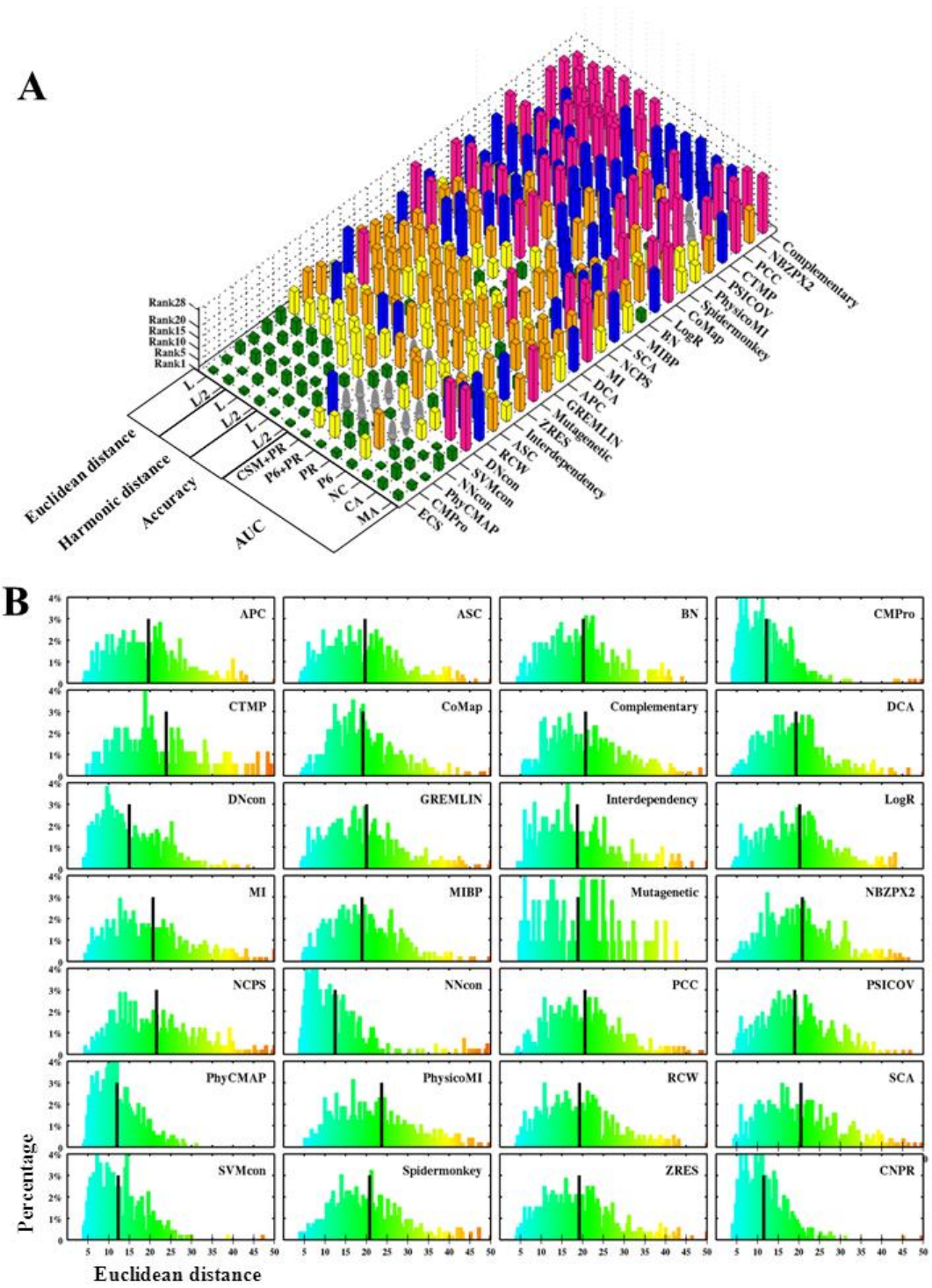


Figure 5.2: Evaluation of sequence-based methods in predicting HIV-1 intra- and inter-protein coevolution

(A) Evaluation of the method combination CNPR and the 27 individual methods applied to the 7 HIV-1 datasets. The x-axis indicates the 28 assessed methods which are ordered according to their performance ranking (CNPR with the best ranking is positioned on the left). The y-axis indicates four statistical measurements (AUC, accuracy, harmonic distance, Euclidean distance) used for the assessment of coevolution predictions given 7 HIV-1 datasets. The L and L/2 top-ranked predictions are evaluated by the measurements of accuracy, harmonic distance and Euclidean

distance. The z-axis indicates the performance ranking of individual methods, where one method with the best ranking has the shortest bar. Based on the performance ranking, bars are colored by green (ranking: 1-5), yellow (ranking: 6-10), orange (ranking: 11-15), blue (ranking: 16-20) and red (ranking: 21-28). Grey cones are used when long-range couplings were not predicted by the corresponding sequence-based methods (e.g. AUC is not evaluated for NNcon in predicting long-range couplings in the p6 protein). Ranking data is provided in **Table S 5.2**.
(B) Distribution plots of Euclidean distance between position pairs in the L top-ranked couplings predicted by individual methods. X- and y-axes indicate the estimated Euclidean distances and the percentage of top-ranked couplings, respectively. Black lines indicate the mean values of Euclidean distances calculated using the L top-ranked couplings. For any method, a lower value of average Euclidean distance indicates that predicted coevolving pairs are in proximity, showing a better prediction performance.

Sequence-based methods cluster according to their methodology

We hypothesized that methods designed from a similar underlying methodology may output similar predictions. To measure the prediction similarities between the sequence-based methods, we calculated Jaccard and association coefficients for the top-ranked predictions between every two methods applied to the 7 HIV-1 datasets. CNPR shared the highest Jaccard and association coefficients with CMPro and PhyCMAP among the 27 sequence-based methods (**Figure 5.3A**). This observation was independent of the prediction cutoffs (**Figure 5.4**). Our hierarchical clustering analysis on the Jaccard and association coefficients revealed four clusters, each of which contained methods generating similar predictions (**Figure 5.3B**). Among the four methods integrated in CNPR, CMPro and PhyCMAP shared the same cluster with CNPR, while NCPS and RCW were individually located in the other two clusters (**Figure 5.3C**). Moreover, 15 out of 19 methods grouped in the method network were designed using similar methodologies, indicating that methods designed from a similar methodology tend to generate similar predictions.

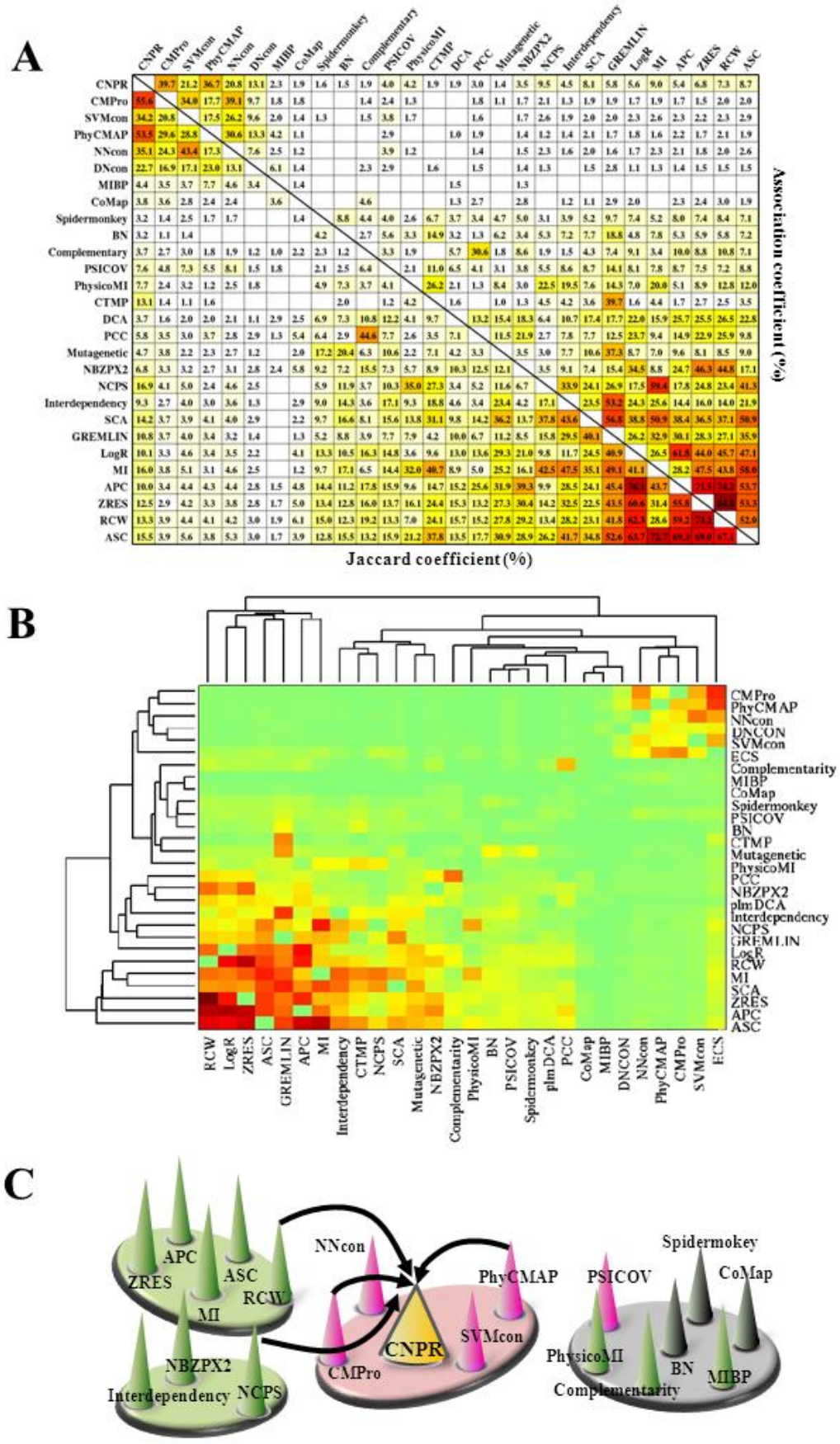


Figure 5.3: Prediction similarity of sequence-based methods and method clustering

(A) Jaccard and association coefficients between the L top-ranked couplings predicted by 28 sequence-based methods.
(B) Hierarchical clustering analysis of Jaccard (bottom) and association (left) coefficients between the 28 sequence-based methods. The heat-map distinguishes the smallest (green) and highest (red) coefficients between the 28 sequence-based methods.
(C) Four method clusters identified commonly by the two clustering trees in (B). The arrows connect four methods (CMPro, NCPS, PhyCMAP, RCW) integrated in CNPR. Methods designed based on mutual information are colored in green, phylogenetics in grey, machine learning in pink.

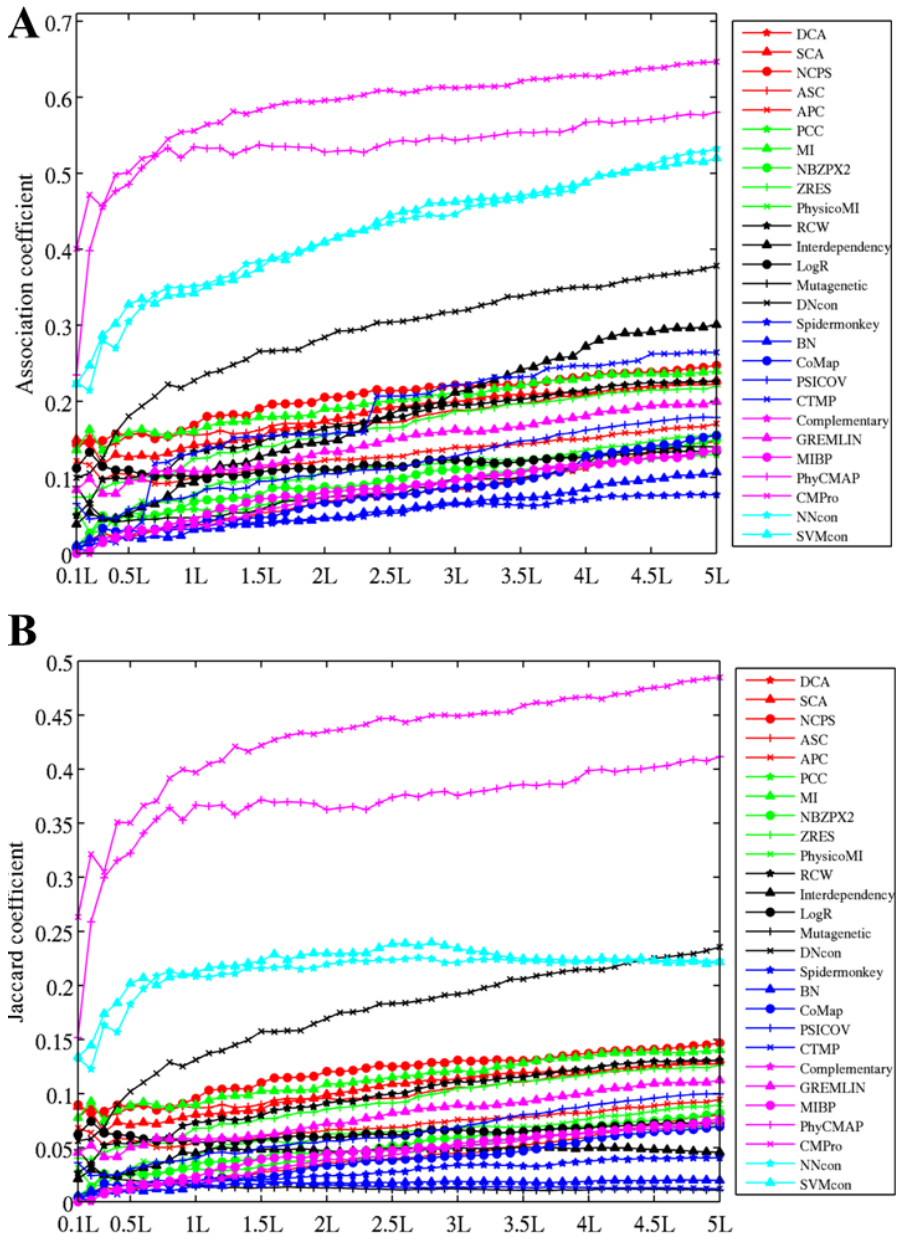


Figure 5.4: Jaccard and association coefficients between CNPR and 27 sequence-based methods. The x-axis indicates the cutoff of the top-ranked couplings used for the calculation of Jaccard and association coefficients, where L is the number of AA sequence in the sequence inputs.

Detection of HIV-1 intra-protein coevolution

Using HIV-1 sequence datasets, we applied CNPR to investigate coevolution within each HIV-1 protein. In this section, the predicted coevolving residues refer to the L top-ranked long-range couplings predicted by CNPR.

Of the 132 predicted coevolving residues in the HIV-1 matrix protein (L = 132), 30.3% were true positives (thus accuracy equals 30.3%), 56.8% were between two helix structures (helix-to-helix), 40.9% involved one position in the third (positions: 47-67) and 50.1% one position in the fourth (positions: 73-90) helix structures (**Figure 5.5**A). The average Euclidean distance of the predicted coevolving residues was 9.97Å compared to 19.22Å between all residue pairs. As an example, CNPR predicted a true positive coupling A45+E74 (Euclidean distance: 5.69Å) within the inter-domain interaction interfaces involving with the third and the fourth random-coil structures in the matrix protein (**Figure 5.5**B).

Of the 231 predicted coevolving residues in capsid (L=231), 21.2% were true positives, 9.5% were between two random-coil structures (coil-to-coil) and 52.8% were helix-to-helix couplings involving heavily 4 of the 11 helices (helix 3: 16.9%, helix 7: 15.2%, helix 11: 19.1%, helix 12: 18.6%) (**Figure 5.5**C). Average Euclidean distance of the predicted coevolving residues was 12.78Å compared to 26.07 Å between all residue pairs. CNPR also predicted the capsid coupling S41+T54 (7.22Å) within the inter-domain interaction interfaces located between N-terminal domains (NTDs) (**Figure 5.5**D).

Of the 99 predicted coevolving residues in protease (L=99), 44.4% were true positives, 79.8% were between two beta-strands (strand-to-strand), 6.1% were coil-to-coil couplings. Many predicted coevolving residues involved one position in the fourth (25.3%), the fifth (52.5%) and the sixth beta-strands (44.4%) (**Figure 5.7**). Average Euclidean distance of the predicted coevolving residues was 9.87Å compared to 17.61Å between all pairwise residues. CNPR did not detect inter-domain couplings between two monomers in protease.

Regarding the coevolution predictions in nucleocapsid (L=52) and p6 (L=55), 100% and 67.05% were in the random-coil structures, respectively. No inter-domain couplings were detected since both nucleocapsid and p6 are monomers.

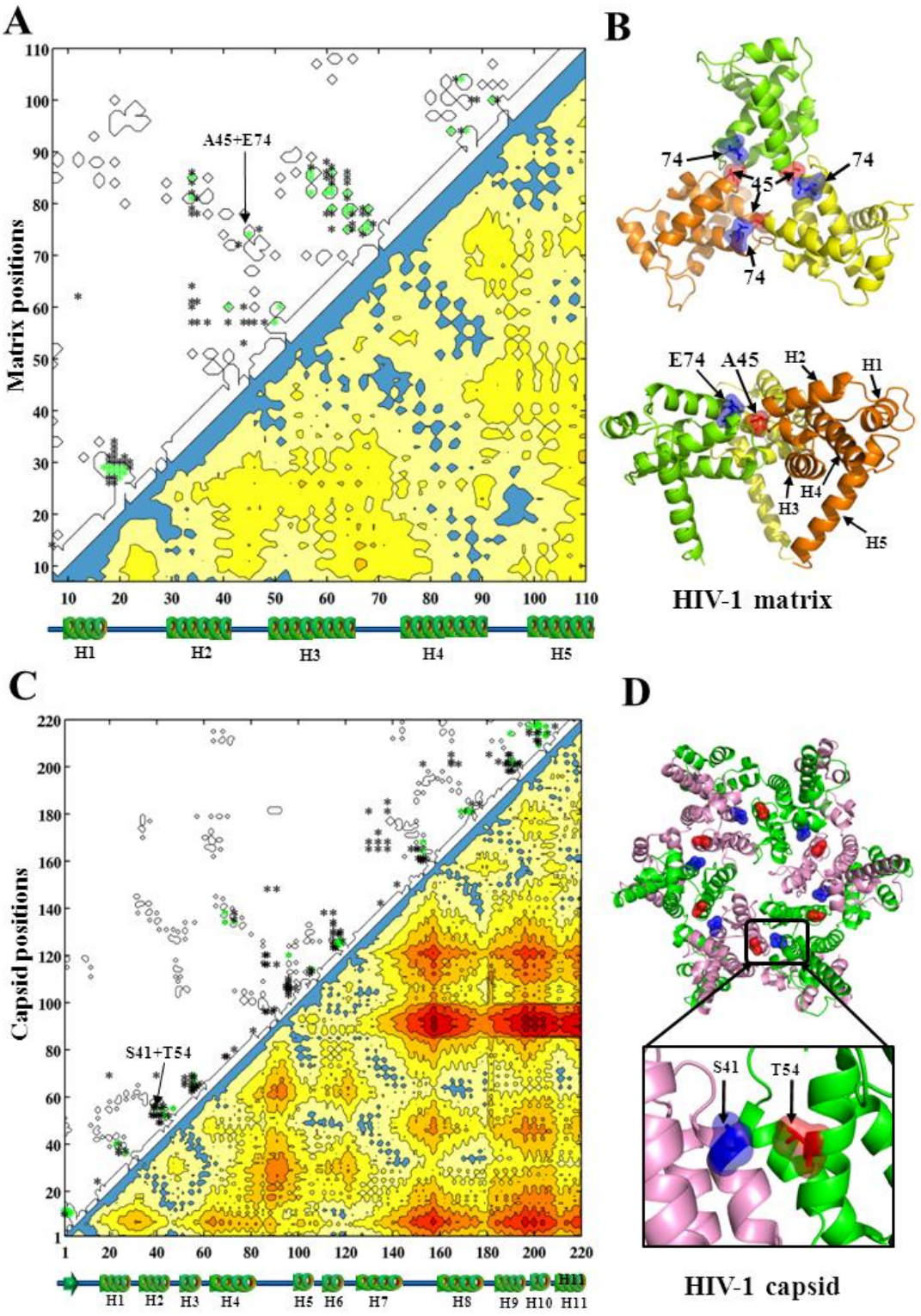


Figure 5.5: Intra-protein couplings of HIV-1 matrix and capsid predicted by CNPR

(A) Contact map of HIV-1 matrix protein and intra-protein coevolving pairs predicted by CNPR. Five helices (H1 to H5) and random-coil secondary structures are aligned to the x-axis. At the bottom right, protein contact map is colored according to the Euclidean distances between two amino acid positions in the 3D structure. Coevolving pairs are colored blue if Euclidean distances were less than 8Å, otherwise gradient from yellow to red. At the upper left, the predicted coevolving residues are marked as asterisks. Green asterisks indicate true positive couplings falling within the black contours of protein contact map.
(B) Cartoon representation of HIV-1 matrix structure. The predicted intra-domain coupling between the residues A45 and E70 is annotated. PDB code: 1HIW.
(C) Contact map of HIV-1 capsid protein and intra-protein coevolving pairs predicted by CNPR. Figure captions are the same as in (A).
(D) Cartoon representation of HIV-1 capsid hexamer with 6 identical units. The predicted intra-domain coupling between the residues A42 and T54 is annotated. PDB code: 3H4E.
The intra-protein couplings predicted by all 28 methods in HIV-1 proteins are shown in **Figure S 5.1**-**Figure S 5.4**. Visualization: PyMOL V1.5(http://www.pymol.org/).

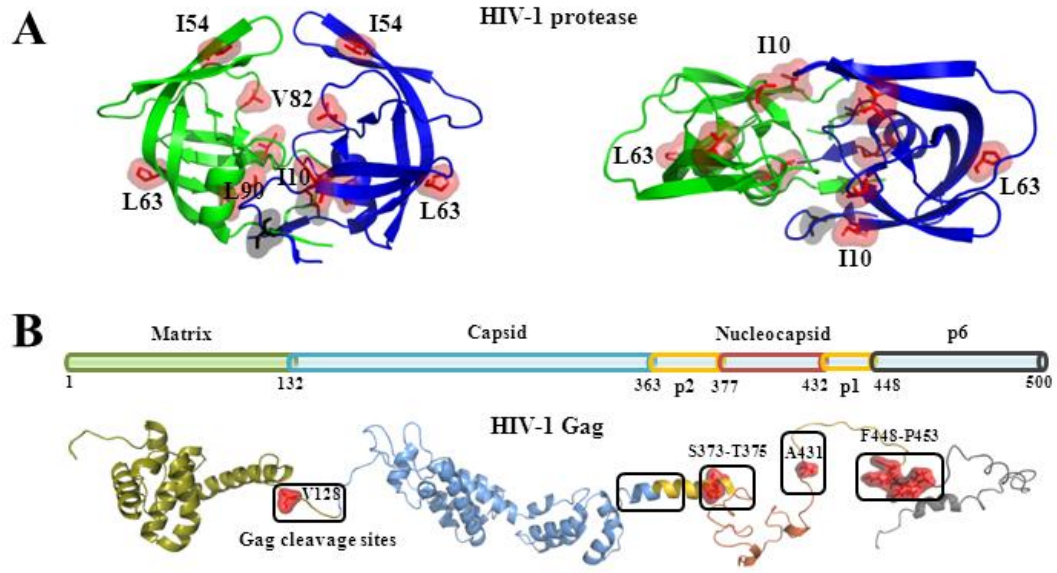


Figure 5.6: Structural representation of gag and protease proteins
(A) Top and side views of the residue positions (T4, L10, I54, L63, V8, L90) in HIV-1 protease. (B) Gag cleavage sites in the 3D protein structure of gag proteins. Gag cleavage sites are annotated in boxes and amino acid positions (V128, S373-T375, V431, F448-P453) are colored in red. PDB code: 1HIW (matrix), 3NTE (capsid), 1U57 (p2), 2M3Z (nucleocapsid),2C55(p6). Visualization software: PyMOL V1.5 (http://www.pymol.org/).

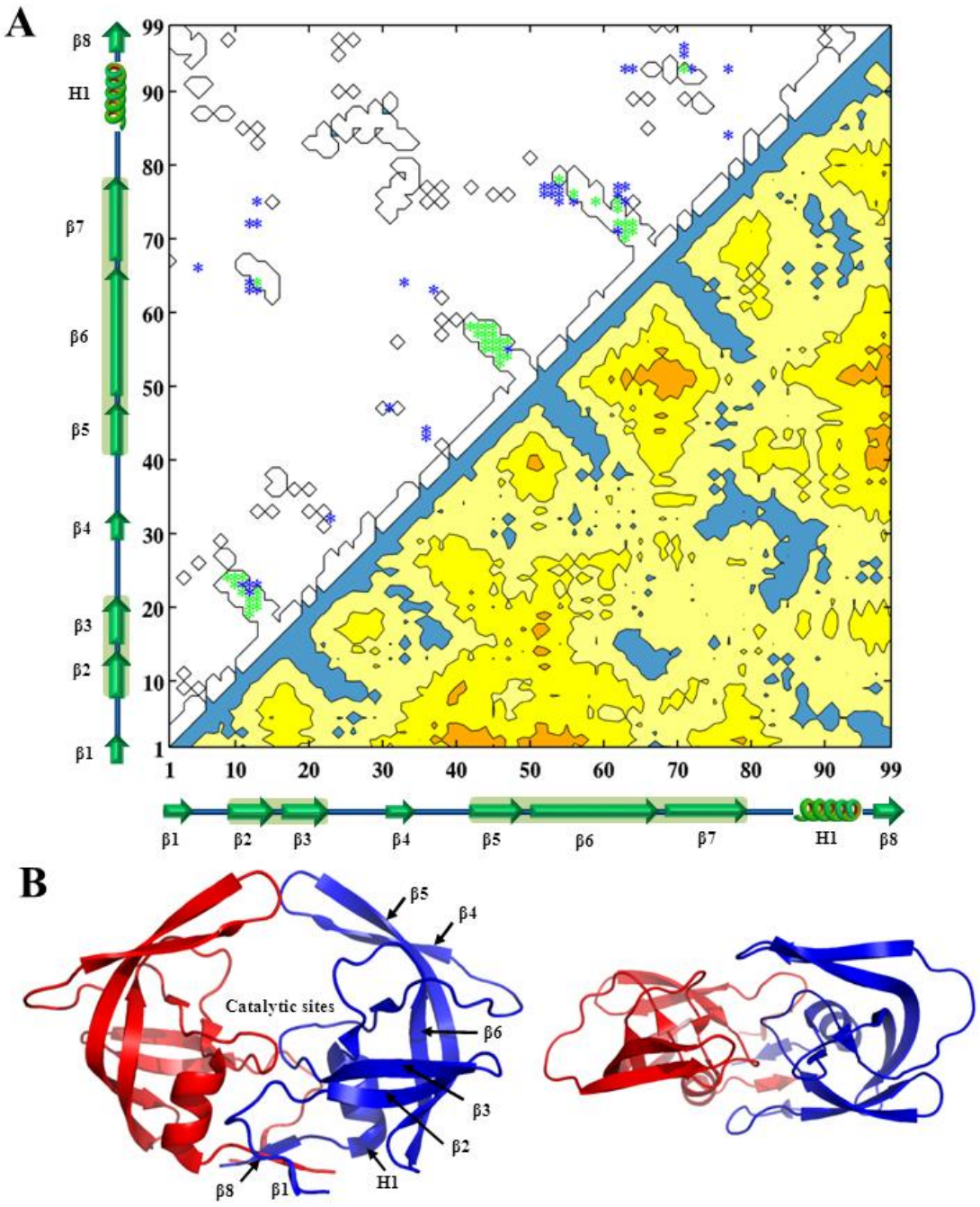


Figure 5.7: HIV-1 protease coevolving pairs predicted by CNPR. (a) The contact map of HIV-1 protease (bottom right) and the predicted coevolving pairs (top left) are illustrated. Green dots indicate true positive predictions in the protein contact map. The random-coil (e.g. L1-L2), beta-strand (e.g. β1-β3) and helix (e.g. H1) secondary structures are shown along the x- and y-axes. (b) The top (right) and side (left) view of HIV-1 protease structure with 2 symmetrical units colored blue and red, respectively. PDB code: 1TW7. Visualization software: Matlab and PyMOL V1.5.

Detection of HIV-1 inter-protein coevolution

We applied CNPR to investigate HIV-1 inter-protein coevolution using the protease-p6 and protease-GCS sequence datasets. In this section, the predicted coevolving

residues refer to the L top-ranked long-range couplings predicted by CNPR. Of the 151 predicted protease-p6 couplings (L=151), 17.9% were true positives, 21.8% were located in the coil-to-coil couplings, 53.3% were coil-to-strand couplings, 28.5% involved 5 protease positions (T4, L10, L63, V82, L90), 76.2% involved either protease cleavage sites Q450-P453 or protease-p6 overlapping positions (Gag positions: S489-Q500) (**Figure 5.6**), 58.9% had either Gag or protease positions identified in experimental studies (**Table S 5.1**).

Of the 149 coevolving residues predicted between protease and GCS (L=149), 28.9% were true positives, 84.6% had either Gag or protease positions identified in the experimental and clinical studies, 25.5% were coil-to-coil couplings, 68.5% were the coil-to-strand couplings, 25.5% involved 4 protease positions (L10, I54, L63, V82), 93.3% had GCS positions V128, S373-T375, A431 and F448-P453. Of interest, protease positions L10, I54, L63 and V82 are located near the protease active site (**Figure 5.6**).

5.5 Discussion

To our knowledge, this study presents the first ensemble coevolution system (ECS) to predict the position-specific coevolution in HIV-1 proteins. Ensemble systems with robust predictions have been applied previously [29-33, 68-71]. For instance, a super learner was created to improve the prediction of HIV-1 drug susceptibility using a set of machine learning algorithms [68]. As shown in our study, an ensemble approach can provide robust predictions of position-specific coevolution when different sequence-based methods predict different coevolving residues. The problem of discordant predictions has been reported previously. For instance, a significant variability in the performance of 13 sequence-based methods was shown using simulated and experimental MSAs [46]. A review which summarized the performance of 9 sequence-based methods also demonstrated different predictions of sequence-based methods [24]. The aim of our study was to detect HIV-1 intra- and inter-protein coevolution using the ensemble learning strategy. For this reason, our study presents a new ensemble coevolution system that integrates 27 sequence-based methods published in the last decade.

An ensemble approach outperforms individual sequence-based methods in detecting HIV-1 coevolution

Armed with our coevolution system, HIV-1 coevolving residues were predicted and the true positive predictions were evaluated using independent evaluation datasets. For HIV-1 intra-protein coevolution, we used protein contact maps to evaluate coevolving residues in close proximity within protein structures. For HIV-1 inter-protein coevolution, we evaluated protease-GCS and protease-p6 couplings using the results reported in literature, summarized in our experimental and clinical datasets (**Table S 5.1**).

We designed a heuristic algorithm to identify CNPR – a combination of four methods (CMPro [39], NCPS [48], PhyCMAP [51], RCW [52]). We found that CNPR outperformed any of the 27 individual methods in the prediction of HIV-1 intra- and inter-protein coevolution. Moreover, CNPR was mostly ranked first or second using four measurements (AUC, accuracy, harmonic distance, Euclidean distance) for performance evaluation (**Table S 5.2**). Interestingly, our clustering analysis showed that the four methods in CNPR originated from three method clusters (**Figure 5.3C**), suggesting that combining methods designed from different principles may establish a superior ensemble method [65]. This observation was supported by a recent study, showing that the combination of PSICOV and plmDCA can improve the prediction performance of either PSICOV or plmDCA alone [72]. Our heuristic algorithm used weighted voting as a combination strategy. During the design of our algorithm, we examined two other ensemble strategies, namely majority voting (predictions supported by more than 50% of the considered methods) and Borda count (predictions made by all the methods) [28], both of which yet failed to outperform individual methods (average rankings beyond the top 10, data not shown). Other advanced ensemble algorithms may provide alternative strategies with promising performance.

Our study aimed at comparing sequence-based methods as accurately as possible, but five factors may limit our comparisons: (1) protein contact maps obtained from crystallized structures may reveal most but not all coevolving residues. The contact map evaluation assumes that a destabilizing mutation at one position is compensated for a mutation at the other position in contact [87], probably due to biochemical constrains (i.e. charge, volume and polarity) [94]. Yet, two residues that are in close

contact may not always coevolve [73, 74]. Coevolving residues are not necessarily in physical contact due to protein dynamics in various contexts [16, 20, 75]. Despite these, protein contact maps remain the most popular strategy to evaluate true positive predictions in position-specific coevolution [22]. (2) Default parameters of sequence-based methods were mostly applied in our study but the optimization of parameters adapted to the HIV-1 datasets may provide better predictions. For instance, phylogenetic methods usually require high computation and memory consumption, forcing less optimized parameters to be used [22]. (3) Experimental and clinical studies provide some but not complete data to evaluate all true positive predictions. (4) The power of position-specific methods relies on the number of mutations observed in MSA inputs, limiting the prediction of coevolution occurring at highly conserved residues [76]. (5) Besides the above factors, phylogenetic bias, indirect coupling and stochastic effects can affect coevolution prediction [46, 50].

HIV-1 intra-protein coevolution detected by the method combination CNPR

We applied the method combination CNPR to investigate HIV-1 intra-protein coevolution in Gag and protease proteins, which play important roles in HIV-1 morphogenesis [1]. While CNPR was selected because it had the highest number of true positive predictions, we also found other interesting observations among the predicted co-evolving residues.

In our analysis of matrix intra-protein coevolution, 30.3% of the predicted coevolving residues were true positives – a promising accuracy which represents a three-fold enrichment compared to a random prediction (average percentage of residue pairs in contact: 10.5%, see **Table 5.3**). Most predicted coevolving residues were located between the third (positions: 47-67) and the other helices in matrix, suggesting a role of the third helix in viral assembly. Previously, positions 54 and 68 were found to be important for matrix assembly [77]. Many positive predictions had residue positions between 45-47 and 68-74 (e.g. A45+E74), formed as two short random-coil loops in the matrix protein. As illustrated in **Figure 5.5**B, these two loops are in contact and located in the inter-domain interaction interface of the matrix trimeric complex. Matrix mutations near this interaction interface can alter the intra-domain interactions, resulting in the impairment of viral assembly and Env incorporation [78, 79].

Table 5.3: Summary of long-range residue contacts derived from HIV-1 Gag and protease protein structures

Protein	Number of intra-domain long-range contact (1)				Number of inter-domain long-range contact (2)				Percentage(3)
	helix-helix	strand-strand	helix-strand	others(4)	helix-helix	strand-strand	helix-strand	others	
Matrix	346	0	0	277	0	0	4	23	650/5995=10.46%
Capsid (5)	520	15	22	680	46	0	19	8	1310/22993=5.59%
Nucleocapsid	4	0	0	192	0	0	0	0	196/1225=14.78%
p6	7	0	0	54	0	0	0	0	61/1081=5.19%
Protease	3	233	42	248	0	0	13	8	547/4371=12.00%
Total	880	248	64	1451	46	0	36	39	2764/36712=7.53%

Long-range residues are defined as two residues have at least 6 amino acids apart in the protein sequence.Two residues are in contacts if the Euclidean distance of their Cα atoms is less than 8 Å in the protein 3D structure.
(1) Intra-domain long-range contacts: the number of long-range contacts of residue pairs (see definitions in Methods) within a protein domain, which are classified according to the type of secondary structures involved (e.g. a helix-strand contact indicates contact between a protein residue in an alpha-helix structure and a protein residue in a beta strand structure).
(2) Inter-domain contacts: residue contacts between different protein domains.
(3) Percentage: the proportion of long-range residue pairs in contact, calculated using PDB data (e.g. for the matrix protein, 650/5995=10.46% indicates that 5995 possible long-range pairs of positions are resolved in the crystal structure and 650 of them are in direct contact).
(4) Others: all the other residue contacts (helix-to-coil, strand-to-coil and coil-to-coil contacts)
(5) Capsid contact map is based on crystalized hexamer.
PDB code: 1HIW (matrix), 3H4E (capsid), 1A1T (nucleocapsid), 2C55 (p6), 1TW7 (protease).

In our analysis of capsid intra-protein coevolution, 21.2% of the predicted coevolving residues were true positives – a four-fold enrichment compared to a random prediction (the average percentage of residue pairs in contact: 5.6%, **Table 5.3**). Half (52.8%) of the long-range coevolving residues were found within helices, especially the helices 3, 7, 11 and 12 (**Figure 5.5**D). These helices near the capsid intra- and inter-domain interaction interfaces play a key role in the capsid assembly and stability [1, 80-82]. The helices 3, 4 and 7 in the N-terminal domain (NTD) and helices 8 and 11 in the C-terminal domain (CTD) are essential for NTD-CTD interactions in the capsid hexamer [81-84]. When considering predicted intra-domain coevolving

residues in capsid, E71+L111 was previously predicted using a dataset of HIV-1, HIV-2 and SIV sequences [6]. In our analysis using CNPR, the predicted coupling S41+T54 was ranked higher than E71+L111. Moreover, the Euclidean distance between S41 and T54 (7.22Å) is shorter than that between E71 and L111 (9.85Å).

In our analysis of protease intra-protein coevolution, 44.4% of the predicted coevolving residues were true positives – a four-fold enrichment compared to a random prediction (the average percentage of residue pairs in contact: 12%, **Table 5.3**). Most statistical couplings (79.8%) were between beta-strand structures; particularly, the second, third and fifth beta-strand structures. Coevolving residue clusters in these beta-strand structures have been reported previously [85, 86].

Besides the intra-protein coevolution reported here, other coevolution events in HIV-1 Gag have also been reported. For instance, five groups of Gag positions were coevolving under multidimensional constraints and one of these groups contains positions in the capsid N-terminal helices [7]. Our coevolution analysis on HIV-1 capsid also identified statistical couplings at the N-terminal helices near the inter-domain interaction interface. In another study, phylogenetic dependency networks were used to infer patterns between human leukocyte antigen (HLA) alleles and HIV-1 Gag residues, resulting in the prediction of 149 couplings between HLA alleles and Gag codons, as well as 1386 couplings within matrix and capsid [5]. Our study observed different predictions within matrix and capsid, possibly because we focused on HIV-1 subtype B, while the coevolution analysis in [5] used a mixed subtype B and C dataset. Further investigation needs to distinguish coevolving residues in HIV-1 subtypes B and C.

HIV-1 inter-protein coevolution estimated by the method combination CNPR

We applied the method combination CNPR to investigate HIV-1 inter-protein coevolution. It is known that the open reading frame of p6 (nucleotides: 120-159) overlaps with protease (nucleotides: 1-40) in the viral genome and that Gag cleavage sites (GCS) interact with protease during the protease-mediated proteolytic processing [4, 87]. Since Gag cleavage sites interact with protease residues, mutations near Gag cleavage sites can be selected under the selective pressure of protease inhibitors [4, 88]. CNPR predicted Gag cleavage sites 128, 373-375, 431 and 448-453 coevolving

with protease residues close to the active site. This is in agreement with previous findings that amino acid substitutions at these Gag cleavage sites are associated with PI resistance [4, 88].

In our analysis of p6-protease inter-protein coevolution, 17.9% of the predicted coevolving residues were true positives and 58.9% contained either a Gag or a protease position in HIV-1 clinical and experimental datasets. In the p6-protease overlapping region (Gag position: 487-500, protease position: 1-13. e.g. T4 and L10), many p6 residues (75.7%) were coupled with protease residues (e.g. T4), illustrating the HIV-1 coevolution in the p6-protease overlapping region. Moreover, p6 residues are mostly coupled with the protease position T4 and protease positions (L10, V82, L90) near the protease substrate-binding pocket (**Figure 5.5**A). Recognized by the known drug resistance algorithms (e.g. IAS-USA, HIVdb, Rega) [89], all these protease positions are associated with PI drug resistance.

Besides the protease-p6 and protease-GCS coevolution, other inter-protein relationships have been reported between Gag proteins. A recent study showed that the p6 residue S40 can partially rescue the negative effects of capsid mutants at the positions E207, A208 and P231 [90]. Matrix can fold back onto nucleocapsid to regulate Gag assembly by the lateral Gag-Gag inter-protein interaction [91]. While the matrix-nucleocapsid interaction interface remains unclear, residues between the matrix domain (positions: 114-126) were coupled with the nucleocapsid domain (positions: 379-383) in our prediction model. Since the predicted coevolving residues do not necessarily imply the spatial proximity or direct protein-protein interactions [24], structural experiments are still needed to clarify the matrix-nucleocapsid interaction domains.

Limitations and future perspectives

Our ensemble approach has its limitations. (1) ECS assembles individual methods so that combinations of methods cannot reveal coevolving residues that are absent in the predictions of individual methods. (2) For some datasets, the method combination CNPR does not always perform the best compared to individual methods. However, it does provide robust predictions with the highest overall ranking in our performance evaluation (**Table S 5.2**). (3) It can be computationally expensive to assemble

prediction results obtained from multiple methods, especially when phylogeny-based methods are integrated. According to our experience, it usually takes more than 30 hours to test a single dataset using all 27 methods (system settings: Linux, CPU 2.8GHz×4). High-standard file management is also needed to organize different inputs and outputs for the 27 methods.

Our study aimed at detecting coevolution in different HIV-1 proteins and our performance comparison was restricted to HIV-1 datasets. Future analysis is still needed to improve the computation efficiency of ECS and to examine the performance of ensemble methods using large-scale protein family datasets. As new sequence-based coevolution methods continue to be reported [22], future studies also need to integrate new methods in the ensemble coevolution system.

5.6 Conclusions

Our study presents a new ensemble coevolution system that integrates multiple sequence-based methods to improve the prediction of HIV-1 position-specific coevolution. Using HIV-1 structural and experimental data, this ensemble system enabled us to identify a combination of 4 different methods that outperformed 27 sequence-based methods for the prediction of HIV-1 inter- and intra-protein coevolution. We also investigated HIV-1 intra- and inter-protein coevolution by exploring coevolving residues in the HIV-1 Gag and protease proteins, which are responsible for virion morphogenesis. Overall, our ensemble coevolution system can detect HIV-1 intra- and inter-protein coevolution, leading to a better understanding of coevolution at the molecular level.

5.7 Additional file 1: Figures

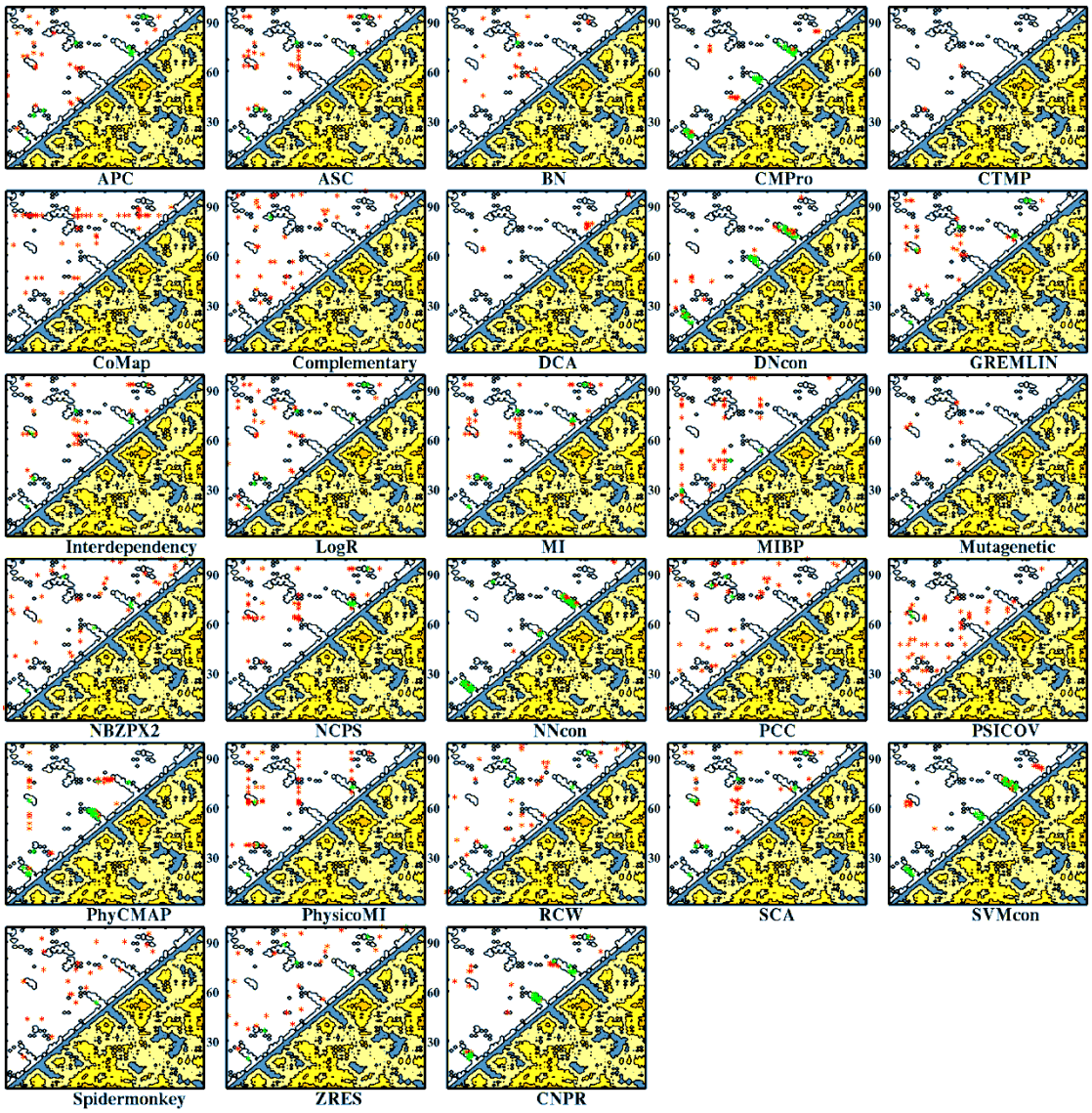


Figure S 5.1: Contact map of HIV-1 protease and coevolving pairs predicted by 28 sequence-based methods. For each subplot, the protein contact map is shown at the bottom right and the top-ranked coevolving residues (L=99) predicted by sequence-based methods are shown as asterisk in the upper left side. True positive coevolving pairs are those falling within the contours of the protein contact map, and indicated as green asterisk. The others are red.

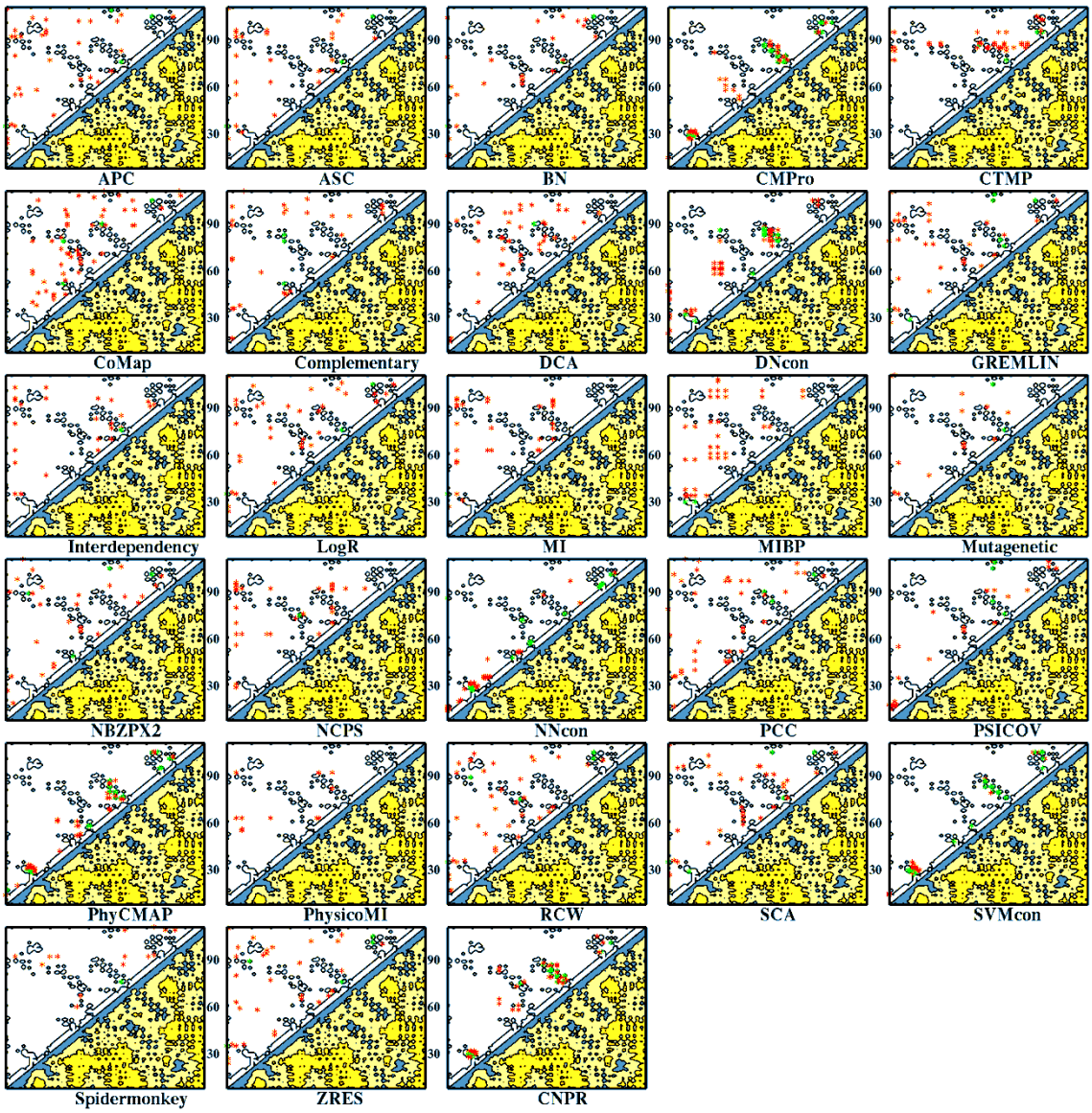


Figure S 5.2: Contact map of HIV-1 matrix and coevolving pairs predicted by 28 sequence-based methods. For each subplot, the protein contact map is shown at the bottom right and the top-ranked coevolving residues (L=99) predicted by sequence-based methods are shown as asterisk in the upper left side. True positive coevolving pairs are those falling within the contours of the protein contact map, and indicated as green asterisk. The others are red.

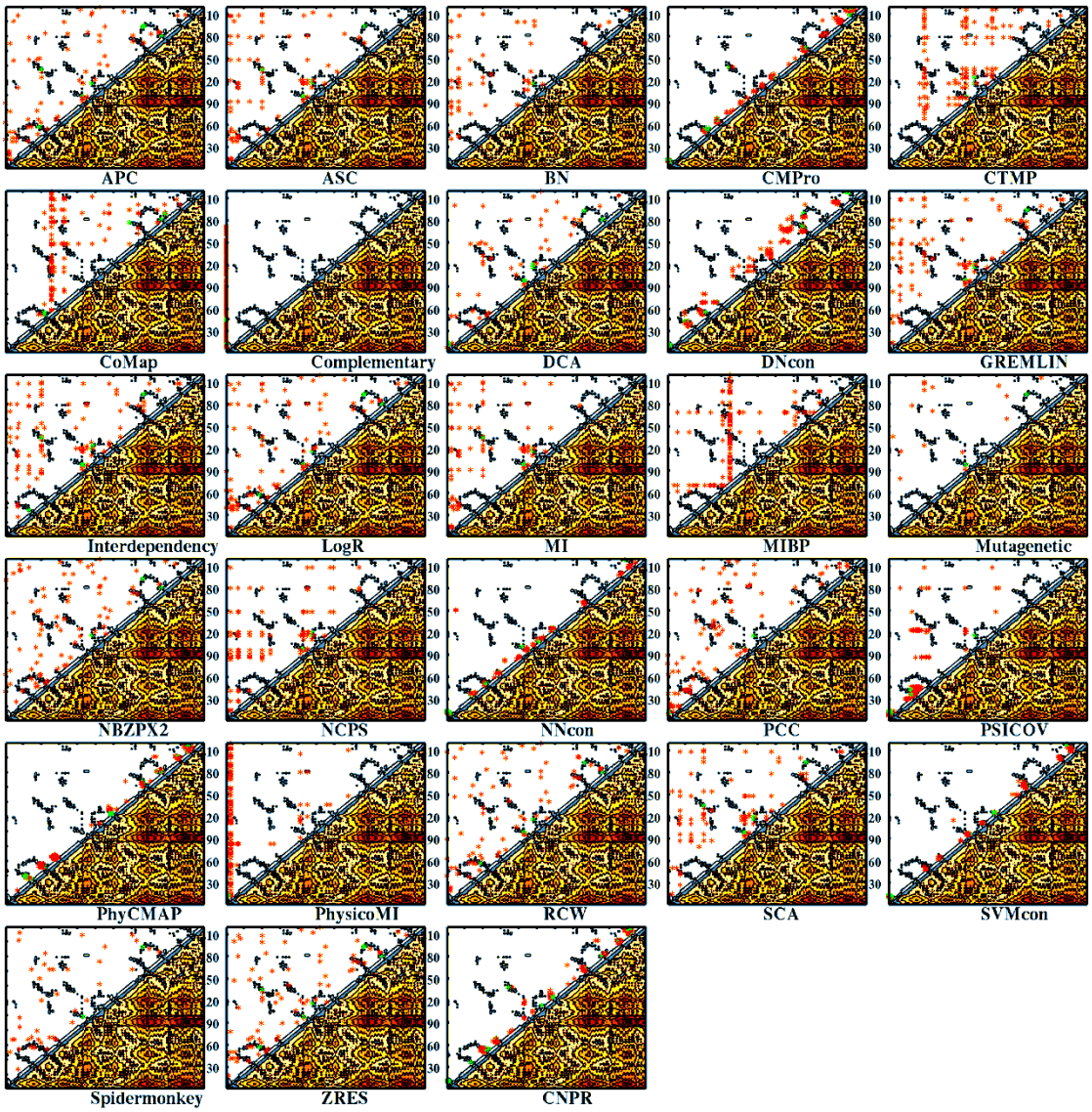


Figure S 5.3: Contact map of HIV-1 capsid and coevolving pairs predicted by 28 sequence-based methods. For each subplot, the protein contact map is shown at the bottom right and the top-ranked coevolving residues (L=231) predicted by sequence-based methods are shown as asterisk in the upper left side. True positive coevolving pairs are those falling within the contours of the protein contact map, and indicated as green asterisk. The others are red.

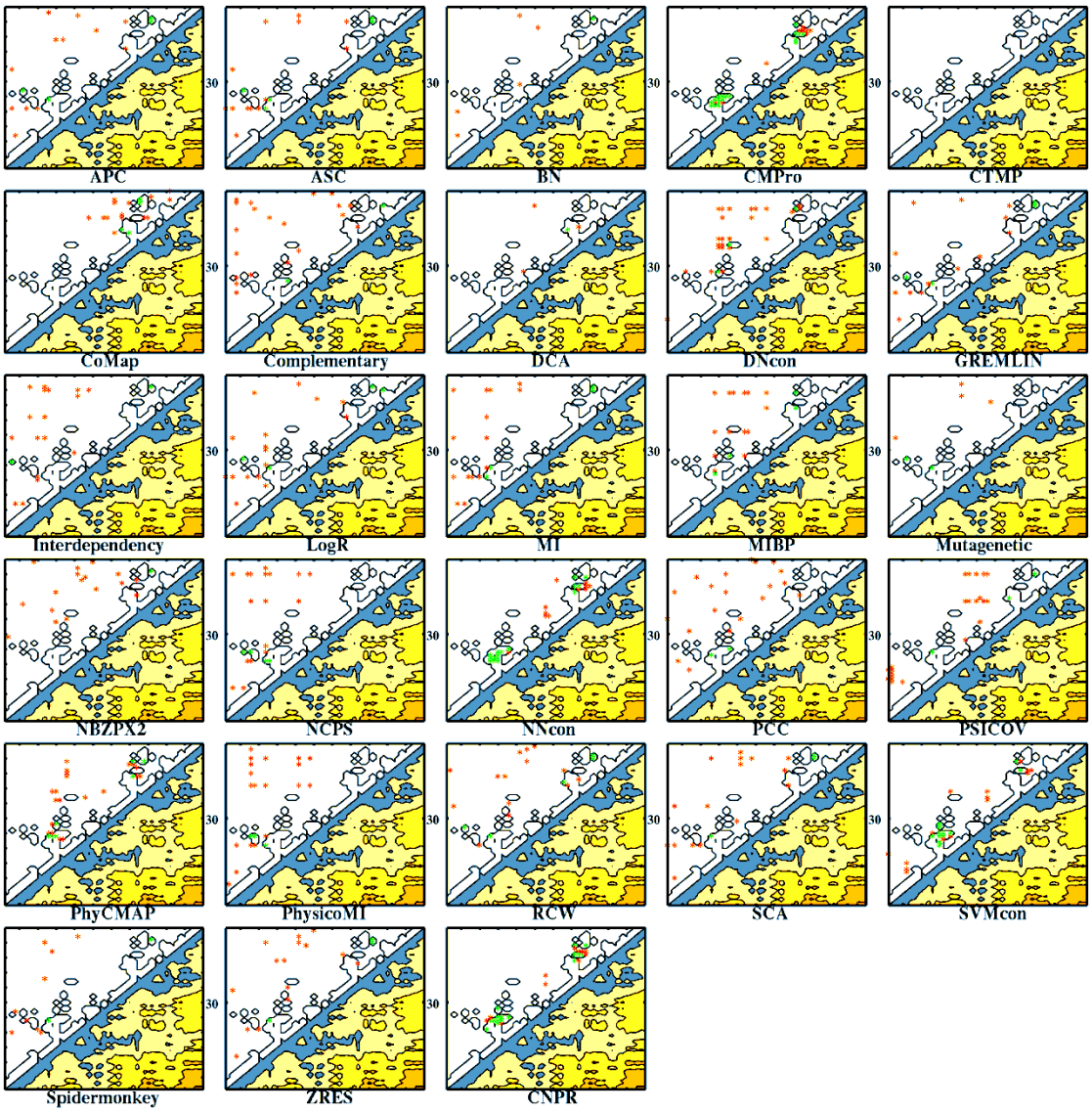


Figure S 5.4: HIV-1 nucleocapsid contact map and coevolving pairs predicted by 28 sequence-based methods. For each subplot, the protein contact map is shown at the bottom right and the top-ranked coevolving residues (L=55) predicted by sequence-based methods are shown as asterisk in the upper left side. True positive coevolving pairs are those falling within the contours of the protein contact map, and indicated as green asterisk. The others are red.

5.8 Additional file 2: Tables

Table S 5.1: Summary of PI-associated Gag and protease substitutions reported in *in vitro* or *in vivo* studies.

Gag substitutions #	Protease substitutions	Ref
L449F+P453T	L10F+G16E+K20T+A28S+M46I+A71V	[92]
R452K	Q58E+A71V	[92]
L449F	L10F	[93]
L449F	L10F+I84V	[93]
L449F	L10F+M46I+I50V	[93]
L449F	L10F+M46I+I47V+I50V	[93]
R452S	L10F+K20I+M36I+M46I+I54V+L63P+A71V+G73S+I84A	[94]
L449F	L10F+M46I+L63P+A71V+I84A	[93, 94]
A431V+L449Q	L10V+M46I+L63A/P+A71V+I84A	[94]
A431V	L10I+M46I+L63P+A71V+L76V+I84A	[94]
A431V+S451I	L10F+L19I+M46I+I47V+I54V+L63P+A71V+I84A	[94]
A431V	L10V+K20I+M36I+M46I+A71V+G73S+L76V+I84A	[94]
L449F	L10I+M46I+L63P+A71V+V77I+I84A	[94]
A431V	L10I+M46I+L63H+A71V+V77I+I84A	[94]
A431V+L449V+R452K	L10F+K20I+M46I+I54M+L63P+A71V+G73T+V77I+I84A	[94]
A431V+L449F	L10F+M46I+L63P+A71V+V77I+I84A	[94]
A431V+L449Q+S451T	L10I+L33F+M46I+I54V+L63P+A71V+L76V+I84A	[94]
R452S	L10F+K20I+M36I+M46I+I54V+L63P+A71V+G73S+I84A	[94]
S451N	K20I+M36I+V82I+I84C	[94]
L449F	L10I+L19I+L24I+L63H+I84C	[94]
A431V	L10I+G16A+M46I+L63P+L76V+I84C	[94]
A431V+L449F	L10F+L19V+L24I+M46L+L63P+I84C	[94]
A431V+L449F	L10I+L24I+M46L+L63P+A71T+G73S+V77I+I84C	[94]
A431V	M46I+L76V	[95]
A431V	M46I	[95]
A431V	L76V	[95]
P453L	D30N+N88D	[96]
E12K+L75R+H219Q+V390D+R409K	L10F+D30N+K45I+A71V	[97]
E12K+L75R+H219Q+V390D+R409K	D30N+M46I+V77I	[97]
L75R+H219Q+V390D	L10F+V32I+M46I+I84V	[98]
A431V	L10F+K20T+L33F+M36I+M46I+I54V+L63P	[99]
A431V	L10I+K20R+L33F+M46L+I54L+L63P+A71V+G73S+V82A+L90M	[99]
I437V	L10I+G48V+I54V+L63P+V77I+V82A	[99]
L483P+K490R	L10I+I15V+E34Q+M36I+T37N+I54A+Q58E+V82A	[100]
I376V+ L483P+K490R	L10I+I15V+E34Q+M36I+T37N+I54A+Q58E+V82A	[100]
I376V+E398V+L483P+K490R	L10I+I15V+E34Q+M36I+T37N+I54A+Q58E+V82A	[100]
L449P+P453L	L19Ins+E21D+A22V+M46I/L63P+A71V+I84V+I93L	[101]
L449P+P453L	M46I+L63P+A71V+I84V+I93L	[101]
L449F	I50V	[102]
L449F	M46I+I50V	[102]
P453L	I50V	[102]
P453L	M46I+I50V	[102]
A431V	L90M	[103]
A431V	N88D+L90M	[103]
A431V	N88D	[103]
A431V	D30N+N88D+L90M	[103]
A431V	D30N+N88D	[103]
A431V	I84V	[103]
A431V	V82A	[103]
A431V	I50L	[103]
A431V	I84V+L90M	[103]
A431V	V82A+L90M	[103]

Gag substitutions #	Protease substitutions	Ref
K436R	L90M	[103]
K436R	I84V+L90M	[103]
K436R	I84V	[103]
K436R	I50V	[103]
I437V	D30N+N88D	[103]
I437V	I50V	[103]
I437V	I84V+L90M	[103]
I437V	V82A	[103]
I437V	V82A+L90M	[103]
I437V	I84V	[103]
L449F	D30N+N88D	[103]
L449F	V82A	[103]
L449F	I50V	[103]
L449F	V82A+L90M	[103]
L449F	L90M	[103]
L449F	N88D	[103]
R452S	I84V+L90M	[103]
R452S	I84V	[103]
R452S	L90M	[103]
P453L	I84V	[103]
P453L	I84V+L90M	[103]
P453L	L90M	[103]
P453L	V82A	[103]
P453L	V82A+L90M	[103]
A431V	M46I +L76V	[104]
P453L	I84V	[105]
A431V	M46I/L,V82A/F/T	[105]
I437A	V82A	[106]
I437V	G48V,I50V,I54A/V,V82A/T	[106]
P459Ins	V82A/F/T/S	[107]
S451N	L10I	[108]
I437T/V	L76V	[109]
A431V	M46I	[110]
N382A	I15V	[111]
A431V	M46L/I+I54V+ V82A	[112]
L449P,S451N,P453L	D30N+N88D	[113]
A431V	L24I+M46I/L+I54V+V82A	[114]
I437V	I54V+V82F/T/S	[114]
L449V	I54M/L/S/T/A	[114]
L449F+R452S+P453L	D30N+I84V	[114]
P453L	V82A	[114]
L449F,S451N/T	D30N+N88D	[115]
A431V	M46I/L,I54V,V82A/T/F	[116]
S373Q,L449P	K20I/R/M,L89M/I	[116]
S125K+Y132F+G62R+I437V	L10I+A71V+N88S	[117]
P453L	N88D	[118]

#: Substitution: the symbol “+” indicates multiple amino acid substitutions observed simultaneously (e.g. Y79F+T81A indicates the presence of both Y79F and T81A).

Table S 5.2: Ranking of sequence-based methods using individual HIV-1 datasets

Sequence-based method	Area-under-curve (AUC)							Accuracy(1)		Harmonic distance(2)		Euclidean distance(3)	
	MA	CA	NC	p6	PR	PR-p6	PR-GCS	L/2	L	L/2	L	L/2	L
APC	11.5*	12.5	14.5	12	14.5	13.5	11	16	15	11.5	12.5	13	14
ASC	15	20	14.5	10	11	12	14	11	8	7	10	7	15
BN	4.5	12.5	10	13	4	18.5	0	24	22	26.5	23	27	24
CMPro	2.5	5.5	1	9	5	10	7	2	3	1	3	1	2
CoMap	25.5	23.5	11	0	19.5	0	22.5	27	27	14.5	9	9	8
Complementary	25.5	23.5	19.5	19	22.5	20	19	26	25	26.5	26	23	23
CTMP	21	23.5	0	0	14.5	9	0	28	28	28	24.5	11	7
DCA	18.5	12.5	14.5	8	26	11	10	17	17	13	17	14	12
DNcon	28	26.5	8	0	9	0	5.5	8	10	6	6	6	6
GREMLIN	15	16.5	12	6	10	16	13	12	14	11.5	15.5	12	16
Interdependency	7	7.5	6	0	7	0	0	19	18	14.5	12.5	18	13
LogR	18.5	16.5	24.5	7	19.5	17	19	15	16	18	18.5	19	21
MI	27	16.5	17.5	2	12.5	4	3	6	7	9.5	12.5	15	18
MIBP	11.5	28	19.5	14.5	22.5	13.5	8	25	22	19	15.5	16	10.5
Mutagenetic	23.5	5.5	5	0	8	2.5	16	9	6	16.5	12.5	24	20
NBZPX2	15	23.5	24.5	17.5	21	22	22.5	23	22	23.5	24.5	26	25.5
NCPS	9.5	26.5	24.5	4	17	2.5	2	7	9	20.5	20.5	25	27
NNcon	6	2	3	0	1	0	0	3	2	3	1.5	3	3
PCC	23.5	9.5	22	0	26	18.5	15	21	24	23.5	27	20.5	25.5
PhyCMAP	1	3.5	4	16	6	5.5	19	5	5	5	5	5	4
PhysicoMI	8	9.5	27	2	28	8	12	20	26	25	28	28	28
PSICOV	15	7.5	24.5	17.5	26	22	21	18	19	22	20.5	20.5	17
RCW	21	20	17.5	5	17	5.5	4	13	11	9.5	7.5	10	10.5
SCA	21	16.5	21	20	12.5	7	5.5	10	12.5	16.5	18.5	17	19
Spidermonkey	9.5	12.5	9	14.5	24	22	17	22	20	20.5	22	22	22
SVMcon	4.5	1	7	0	2.5	0	0	4	4	4	4	4	5
ZRES	15	20	14.5	11	17	15	9	14	12.5	8	7.5	8	9
CNPR	2.5	3.5	2	2	2.5	1	1	1	1	2	1.5	2	1

(1): Rankings are obtained by average accuracy of individual methods (**Table S 5.3).**
(2): Rankings are obtained by average Harmonic distance of individual methods (**Table S 5.4)**
(3): Rankings are obtained by average Euclidean distance of individual methods (

Table S 5.5).
*: the average ranking of the method following the calculation procedures in [119]. We give a simple example to explain the calculation of this average ranking:

	AUC		Ranking		AUC		Ranking
Method 1:	0.9	=>	1	Method 1:	0.8	=>	(1+2)= 1.5
Method 2:	0.8	=>	2	Method 2:	0.8	=>	(1+2)= 1.5
Method 3:	0.7	=>	3	Method 3:	0.7	=>	3

For the example on the right side, the average ranking of the first and second methods is calculated as (1+2)/2 = 1.5. For the method 3, it remains the same ranking as 3.

Table S 5.3: Accuracy of sequence-based methods on individual HIV-1 datasets

Sequence-based method	Accuracy of the L/2 top-ranked long-range couplings								Accuracy of the L top-ranked long-range couplings							
	MA	CA	NC	p6	PR	PR-p6	PR-GCS	Average	MA	CA	NC	p6	PR	PR-p6	PR-GCS	Average
APC	9%	6.9%	17.9%	0%#	22%	9.2%	10.7%	10.8%	9%	4.3%	19.6%	0%	14%	4.6%	8.7%	8.6%
ASC	7.5%	6.9%	21.4%	0%	20%	17.1%	32%	15%	6.8%	4.7%	17.9%	1.9%	19%	11.2%	20.7%	11.7%
BN	6%	2.5%	7.1%	0%	16%	6.6%	2.9%	5.9%	5.6%	2.5%	6.5%	1.9%	11%	6.4%	2.9%	5.2%
CMPro	41.8%	26.7%	64.3%	0%	64%	0%	5.3%	28.9%	30.8%	17.7%	50%	1.9%	49%	0%	8%	22.5%
CoMap	7.5%	5.2%	10.7%	0%	0%	0%	4%	3.9%	6.8%	3.9%	8.9%	0%	4%	0%	6.7%	4.3%
Complementary	4.5%	5.2%	7.1%	0%	6%	5.3%	0%	4%	6%	4.3%	3.6%	0%	6%	5.3%	8%	4.7%
CTMP	9.3%	2.7%	0%	0%	0%	11.1%	0%	3.3%	9.3%	2.7%	0%	0%	0%	11.1%	0%	3.3%
DCA	3%	9.5%	17.9%	3.7%	4%	5.3%	21.3%	9.2%	6%	6%	14.3%	1.9%	4%	3.9%	13.3%	7.1%
DNcon	20.9%	18.1%	10.7%	0%	62%	0%	4%	16.5%	17.3%	14.2%	8.9%	0%	35%	0%	3.7%	11.3%
GREMLIN	10.4%	5.2%	17.9%	0%	26%	13.2%	24%	13.8%	6.8%	3%	12.5%	0%	17%	9.2%	18%	9.5%
Interdependency	4.5%	7.7%	11.1%	4.2%	20%	0%	3.3%	7.3%	5.7%	7.7%	11.1%	4.2%	171%	0%	3.3%	7%
LogR	6%	7.8%	17.9%	0%	18%	9.2%	21.3%	11.4%	6%	5.2%	14.3%	0%	15%	5.3%	12.7%	8.3%
MI	7.5%	5.2%	17.9%	7.4%	24%	23.7%	40%	17.9%	4.5%	3.9%	16.1%	3.8%	16%	18.4%	25.3%	12.6%
MIBP	6%	0%	17.9%	0%	8%	0%	0%	4.5%	5.3%	1.7%	17.9%	0%	12%	0%	0%	5.3%
Mutagenetic	10%	9.1%	20%	0%	22.2%	14.3%	35.7%	15.9%	10%	9.1%	20%	0%	222%	14.3%	35.7%	15.9%
NBZPX2	10.4%	3.4%	7.1%	0%	14%	2.6%	5.3%	6.1%	6.8%	3%	8.9%	0%	10%	2%	6%	5.2%
NCPS	0%	3.4%	17.9%	7.4%	12%	31.6%	46.7%	17%	0.8%	2.2%	10.7%	3.8%	14%	21.1%	28.7%	11.6%
NNcon	37.2%	22.4%	67.9%	0%	70%	2.6%	0%	28.6%	37.2%	20%	44.6%	0%	62%	2.6%	0%	23.8%
PCC	7.5%	3.4%	7.1%	0%	6%	7.9%	17.3%	7%	4.5%	4.3%	3.6%	0%	9%	3.9%	0.1%	5%
PhyCMAP	34.3%	23.3%	28.6%	0%	44%	2.6%	2.7%	19.4%	33.8%	15.9%	30.4%	0%	33%	4.6%	2.7%	17.2%
PhysicoMI	3%	0.9%	10.7%	7.4%	12%	5.3%	10.7%	7.1%	1.5%	0.4%	7.1%	3.8%	7%	5.3%	7.3%	4.6%
PSICOV	10.4%	13.8%	10.7%	0%	4%	11.8%	8%	8.4%	7.5%	6.9%	5.4%	0%	9%	7.9%	6.7%	6.2%
RCW	9%	6.9%	21.4%	0%	14%	13.2%	21.3%	12.3%	8.3%	4.3%	19.6%	1.9%	14%	10.5%	18%	10.9%
SCA	7.5%	5.2%	10.7	0%	20%	28.9%	37.3%	15.7%	6%	3.9%	08.9%	0%	14%	17.8%	24.7%	10.8%
Spidermonkey	4.5%	4.3%	14.3	0%	8%	2.6%	12%	6.5%	3.8%	3.4%	10.7%	0%	9%	2.6%	0.1%	5.7%
SVMcon	47.8%	19.8%	46.4	0%	58%	0%	0%	24.6%	38.5%	18.8%	32.1%	0%	39%	0%	0%	18.3%
ZRES	10.4%	6.9%	17.9	3.7%	16%	7.9%	21.3%	12%	8.3%	4.3%	19.6%	3.8%	13%	7.9%	18%	10.7%
CNPR	38.8%	25%	57.1%	7.4%	56%	19.7%	38.7%	34.7%	30.3%	21.2%	42.9%	3.8%	44%	17.9%	28.9%	26.9%

Table S 5.4: Harmonic distance of sequence-based methods using individual HIV-1 datasets

Sequence-based method	Harmonic distance of the L/2 top-ranked long-range couplings						Harmonic distance of the L top-ranked long-range couplings					
	MA	CA	NC	p6	PR	Average	MA	CA	NC	p6	PR	Average
APC	0.032	0.037	0.043	0.012	0.073	0.039	0.018	0.024	0.034	0.017	0.04	0.027
ASC	0.025	0.029	0.066	0.057	0.078	0.051	0.015	0.02	0.024	0.022	0.058	0.028
BN	0.014	0.004	0	-0.013	0.04	0.009	0.01	0.004	-0.01	0.002	0.035	0.008
CMPro	0.177	0.161	0.222	0.07	0.198	0.166	0.139	0.113	0.175	0.064	0.16	0.13
CoMap	0.033	0.023	0.032	0.078	-0.022	0.029	0.032	0.017	0.027	0.078	-0.007	0.029
Complementary	-0.003	0.034	-0.013	0.005	0.016	0.008	0.003	0.023	-0.025	0.001	0.01	0.003
CTMP	0.045	-0.003	0	-0.038	0.017	0.004	0.045	-0.003	0	-0.038	0.017	0.004
DCA	0.027	0.048	0.047	0.028	-0.003	0.03	0.023	0.037	0.045	0.007	0.002	0.023
DNcon	0.119	0.103	0.059	0.003	0.182	0.093	0.104	0.086	0.054	-0.013	0.119	0.07

GREMLIN	0.034	0.024	0.041	0.015	0.087	0.04	0.019	0.005	0.023	0.009	0.061	0.024
Interdependency	0.016	0.028	-0.022	0.046	0.071	0.028	0.016	0.028	-0.022	0.046	0.062	0.026
LogR	0.012	0.037	0.029	-0.01	0.053	0.024	0.006	0.024	0.016	-0.007	0.037	0.015
MI	0.017	0.017	0.023	0.078	0.08	0.043	-0.002	0.009	0.018	0.045	0.06	0.026
MIBP	0.048	0.009	0.049	-0.016	0.017	0.021	0.041	0.013	0.057	-0.029	0.035	0.023
Mutagenetic	0.029	0.016	0.034	-0.034	0.09	0.027	0.029	0.016	0.034	-0.034	0.09	0.027
NBZPX2	0.031	0.018	-0.004	-0.011	0.023	0.011	0.009	0.017	-0.003	-0.011	0.011	0.005
NCPS	-0.013	-0.005	0.006	0.066	0.037	0.018	-0.018	-0.01	-0.007	0.038	0.053	0.011
NNcon	0.134	0.135	0.197	0.064	0.209	0.148	0.134	0.116	0.146	0.064	0.2	0.132
PCC	0.024	0.027	-0.003	0.008	0.008	0.013	0.001	0.02	-0.023	-0.01	0.011	0
PhyCMAP	0.155	0.159	0.121	0.025	0.131	0.118	0.157	0.135	0.118	0.018	0.108	0.107
PhysicoMI	-0.002	-0.041	-0.016	0.051	0.052	0.009	-0.014	-0.029	-0.01	0.031	0.018	-0.001
PSICOV	0.051	0.063	-0.022	-0.007	-0.003	0.016	0.032	0.036	-0.019	0	0.013	0.012
RCW	0.03	0.042	0.048	0.057	0.042	0.044	0.02	0.024	0.04	0.03	0.043	0.032
SCA	0.032	0.018	0.015	-0.008	0.079	0.027	0.015	0.007	0.004	0.001	0.052	0.016
Spidermonkey	0.009	0.012	0.016	0.011	0.041	0.018	0.001	0.013	0.003	0.004	0.031	0.01
SVMcon	0.192	0.128	0.152	0.037	0.189	0.14	0.165	0.124	0.113	0.027	0.128	0.111
ZRES	0.034	0.044	0.053	0.061	0.041	0.046	0.022	0.024	0.037	0.044	0.035	0.032
CNPR	0.175	0.155	0.19	0.08	0.178	0.155	0.144	0.137	0.16	0.068	0.152	0.132

Table S 5.5: Average Euclidean distance of the top-ranked long-range couplings predicted by sequence-based methods

Sequence-based method	Average Euclidean distance of the L/2 top-ranked long-range couplings (Å)						Average Euclidean distance of the L top-ranked long-range couplings (Å)					
	MA	CA	NC	p6	PR	Average	MA	CA	NC	p6	PR	Average
APC	16.45	21.64	14.74	20.81	13.25	17.38	17.72	22.9	16.69	20.65	15.06	18.6
ASC	16.69	22.84	13.53	16.67	12.31	16.41	17.73	23.54	17.63	20.54	14.01	18.69
BN	18.27	25.53	17.65	23.8	14.46	19.94	18.92	25.53	19.09	22.34	14.76	20.13
CMPro	9.01	11.19	7.47	14.2	8.38	10.05	10.51	14.7	8.8	15.33	9.49	11.77
CoMap	15.63	21.88	14.56	14.31	17.87	16.85	15.95	22.87	15.14	14.31	17.46	17.14
Complementary	19.27	20.32	19.32	20.95	15.52	19.08	18.82	22.6	19.91	22.35	16.37	20.01
CTMP	15.36	27.88	0	27.72	13.91	16.98	15.36	27.88	0	27.72	13.91	16.98
DCA	15.41	20.27	14.53	19.76	17.18	17.43	16.52	21.6	15.18	21.96	17	18.45
DNcon	11.18	15.69	12.55	19.74	9.14	13.66	11.6	16.62	13.87	22.35	11.12	15.11
GREMLIN	16.58	22.21	14.8	19.7	12.39	17.14	17.78	24.97	16.2	21.52	13.37	18.77
Interdependency	17.87	23.22	20.34	17.54	13.02	18.4	18.41	23.22	20.34	17.54	13.38	18.58
LogR	17.68	21.95	15.72	22.64	14.2	18.44	18.24	22.92	16.99	23.08	15.36	19.32
MI	17.66	24.05	16.98	16.73	12.59	17.6	19.17	25.57	17.81	18.79	13.45	18.96
MIBP	14.06	21.77	13.89	23.39	15.9	17.8	14.52	21.93	13.49	25.55	15.13	18.12
Mutagenetic	17.43	23.12	15.53	27.77	11.82	19.13	17.43	23.12	15.53	27.77	11.82	19.13
NBZPX2	16.52	23.51	17.53	23.77	16.23	19.51	18.22	23.68	18.31	24.03	16.76	20.2
NCPS	19.91	26.33	19.29	16.98	14.34	19.37	20.6	27.69	19.87	19.63	13.57	20.27
NNcon	10.87	14.15	8.35	15.26	7.6	11.25	10.87	15.52	10.46	15.26	7.93	12.01
PCC	17.02	21.43	17.38	21.22	16.09	18.63	18.71	22.86	19.26	23.57	16.59	20.2
PhyCMAP	9.58	11.17	10.11	17.6	10.7	11.83	9.55	12.21	10.39	19.01	11.6	12.55
PhysicoMI	18.76	32.18	19.4	18.33	13.65	20.46	20.03	30.06	19.03	20.09	16.08	21.06
PSICOV	14.74	19.37	19.22	22.49	17.33	18.63	16.18	20.36	18.15	22.48	16.77	18.79

RCW	16.71	21.14	15.21	16.67	14.67	16.88	17.36	22.56	15.93	20.07	14.66	18.12
SCA	16.36	24.03	16.07	22.61	12.22	18.26	17.78	24.77	17.39	21.86	13.6	19.08
Spidermonkey	17.81	24.4	16.88	20.56	14.83	18.89	19.06	23.98	18.15	22.24	15.41	19.77
SVMcon	8.52	13.3	9.54	16.82	8.89	11.42	9.48	13.43	11.02	17.96	11.38	12.65
ZRES	16.4	20.75	14.35	16.81	14.94	16.65	17.23	22.69	16.32	18.93	15.24	18.08
CNPR	9	11.23	7.77	13.67	8.65	10.06	10.06	12.78	9.41	15	9.77	11.4

5.9 References

1. Zhao G, Perilla JR, Yufenyuy EL, Meng X, Chen B, Ning J, *et al.* Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* 2013,**497**:643-646.
2. Waheed AA, Freed EO. HIV type 1 Gag as a target for antiviral therapy. *AIDS Res Hum Retroviruses* 2012,**28**:54-75.
3. Bell NM, Lever AM. HIV Gag polyprotein: processing and early viral particle assembly. *Trends Microbiol* 2013,**21**:136-144.
4. Fun A, Wensing AM, Verheyen J, Nijhuis M. Human Immunodeficiency Virus Gag and protease: partners in resistance. *Retrovirology* 2012,**9**:63.
5. Carlson JM, Brumme ZL, Rousseau CM, Brumme CJ, Matthews P, Kadie C, *et al.* Phylogenetic dependency networks: inferring patterns of CTL escape and codon covariation in HIV-1 Gag. *PLoS Comput Biol* 2008,**4**:e1000225.
6. Kalinina OV, Oberwinkler H, Glass B, Krausslich HG, Russell RB, Briggs JA. Computational identification of novel amino-acid interactions in HIV Gag via correlated evolution. *PLoS One* 2012,**7**:e42468.
7. Dahirel V, Shekhar K, Pereyra F, Miura T, Artyomov M, Talsania S, *et al.* Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. *Proc Natl Acad Sci U S A* 2011,**108**:11530-11535.
8. Rhee SY, Liu TF, Holmes SP, Shafer RW. HIV-1 subtype B protease and reverse transcriptase amino acid covariation. *PLoS Comput Biol* 2007,**3**:e87.
9. Rhee SY, Liu TF, Kiuchi M, Zioni R, Gifford RJ, Holmes SP, *et al.* Natural variation of HIV-1 group M integrase: implications for a new class of antiretroviral inhibitors. *Retrovirology* 2008,**5**:74.
10. Beerenwinkel N, Rahnenfuhrer J, Daumer M, Hoffmann D, Kaiser R, Selbig J, *et al.* Learning multiple evolutionary pathways from cross-sectional data. *J Comput Biol* 2005,**12**:584-598.
11. Travers SA, Tully DC, McCormack GP, Fares MA. A study of the coevolutionary patterns operating within the env gene of the HIV-1 group M subtypes. *Mol Biol Evol* 2007,**24**:2787-2801.
12. Bizinoto MC, Yabe S, Leal E, Kishino H, Martins Lde O, de Lima ML, *et al.* Codon pairs of the HIV-1 vif gene correlate with CD4+ T cell count. *BMC Infect Dis* 2013,**13**:173.
13. Theys K, Deforche K, Libin P, Camacho RJ, Van Laethem K, Vandamme AM. Resistance pathways of human immunodeficiency virus type 1 against the combination of zidovudine and lamivudine. *J Gen Virol* 2010,**91**:1898-1908.
14. Fares MA, Travers SA. A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. *Genetics* 2006,**173**:9-23.
15. Lovell SC, Robertson DL. An integrated view of molecular coevolution in protein-protein interactions. *Mol Biol Evol* 2010,**27**:2567-2575.
16. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 1999,**286**:295-299.
17. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 2012,**149**:1607-1621.
18. Ashkenazy H, Kliger Y. Reducing phylogenetic bias in correlated mutation analysis. *Protein Eng Des Sel* 2010,**23**:321-326.
19. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci U S A* 2009,**106**:67-72.
20. Suel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol* 2003,**10**:59-69.

21. Rausell A, Juan D, Pazos F, Valencia A. Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc Natl Acad Sci U S A* 2010,**107**:1995-2000.
22. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. *Nat Rev Genet* 2013,**14**:249-261.
23. Fitch WM, Markowitz E. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem Genet* 1970,**4**:579-593.
24. Horner DS, Pirovano W, Pesole G. Correlated substitution analysis and the prediction of amino acid structural contacts. *Brief Bioinform* 2008,**9**:46-56.
25. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A* 2011,**108**:E1293-1301.
26. Ekeberg M, Lovkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys Rev E Stat Nonlin Soft Matter Phys* 2013,**87**:012707.
27. Liu Y, Bahar I. Sequence evolution correlates with structural dynamics. *Mol Biol Evol* 2012,**29**:2253-2263.
28. Rokach L. Ensemble-based classifiers. *Artificial Intelligence Review* 2010,**33**:1-39.
29. Breiman L. Random forests. *Machine learning* 2001,**45**:5-32.
30. Freund Y, Schapire RE. Experiments with a new boosting algorithm. In: *ICML*; 1996. pp. 148-156.
31. Troć M, Unold O. Self-adaptation of parameters in a learning classifier system ensemble machine. 2010.
32. Gao Y, Huang JZ, Wu L. Learning classifier system ensemble and compact rule set. *Connection Science* 2007,**19**:321-337.
33. Bacardit J, Krasnogor N. Empirical evaluation of ensemble techniques for a pittsburgh learning classifier system. In: *Learning Classifier Systems*: Springer; 2008. pp. 255-268.
34. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 2008,**24**:333-340.
35. Deforche K, Silander T, Camacho R, Grossman Z, Soares MA, Van Laethem K, *et al.* Analysis of HIV-1 pol sequences using Bayesian Networks: implications for drug resistance. *Bioinformatics* 2006,**22**:2975-2979.
36. Yeang CH, Haussler D. Detecting coevolution in and among protein domains. *PLoS Comput Biol* 2007,**3**:e211.
37. Dutheil J, Galtier N. Detecting groups of coevolving positions in a molecule: a clustering approach. *BMC Evol Biol* 2007,**7**:242.
38. Halperin I, Wolfson H, Nussinov R. Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. *Proteins* 2006,**63**:832-845.
39. Di Lena P, Nagata K, Baldi P. Deep architectures for protein contact map prediction. *Bioinformatics* 2012,**28**:2449-2457.
40. Eickholt J, Cheng J. Predicting protein residue-residue contacts using deep networks and boosting. *Bioinformatics* 2012,**28**:3066-3072.
41. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci U S A* 2013,**110**:15674-15679.
42. Tillier ER, Lui TW. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics* 2003,**19**:750-755.
43. Burger L, van Nimwegen E. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput Biol* 2010,**6**:e1000633.
44. Edwards S. Elements of information theory, 2nd edition. *Information Processing & Management* 2008,**44**:400-401.
45. Bremm S, Schreck T, Boba P, Held S, Hamacher K. Computing and visually analyzing mutual information in molecular co-evolution. *BMC Bioinformatics* 2010,**11**:330.
46. Ackerman SH, Tillier ER, Gatti DL. Accurate simulation and detection of coevolution signals in multiple sequence alignments. *PLoS One* 2012,**7**:e47108.
47. Gao H, Dou Y, Yang J, Wang J. New methods to measure residues coevolution in proteins. *BMC Bioinformatics* 2011,**12**:206.
48. Lee BC, Kim D. A new method for revealing correlated mutations under the structural and functional constraints in proteins. *Bioinformatics* 2009,**25**:2506-2513.

49. Tegge AN, Wang Z, Eickholt J, Cheng J. NNcon: improved protein contact map prediction using 2D-recursive neural networks. *Nucleic Acids Res* 2009,**37**:W515-518.
50. Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 2012,**28**:184-190.
51. Wang Z, Xu J. Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics* 2013,**29**:i266-273.
52. Gouveia-Oliveira R, Pedersen AG. Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. *Algorithms Mol Biol* 2007,**2**:12.
53. Poon AF, Lewis FI, Frost SD, Kosakovsky Pond SL. Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models. *Bioinformatics* 2008,**24**:1949-1950.
54. Halabi N, Rivoire O, Leibler S, Ranganathan R. Protein sectors: evolutionary units of three-dimensional structure. *Cell* 2009,**138**:774-786.
55. Cheng J, Baldi P. Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics* 2007,**8**:113.
56. Little DY, Chen L. Identification of coevolving residues and coevolution potentials emphasizing structure, bond formation and catalytic coordination in protein evolution. *PLoS One* 2009,**4**:e4762.
57. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 2010,**27**:221-224.
58. Li G, Verheyen J, Rhee SY, Voet A, Vandamme AM, Theys K. Functional conservation of HIV-1 gag: implications for rational drug design. *Retrovirology* 2013,**10**:126.
59. Minh BQ, Vinh le S, von Haeseler A, Schmidt HA. pIQPNNI: parallel reconstruction of large maximum likelihood phylogenies. *Bioinformatics* 2005,**21**:3794-3796.
60. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 2006,**22**:2688-2690.
61. Hooft RW, Vriend G, Sander C, Abola EE. Errors in protein structures. *Nature* 1996,**381**:272.
62. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The binormal assumption on precision-recall curves. In: *Pattern Recognition (ICPR), 2010 20th International Conference on*: IEEE; 2010. pp. 4263-4266.
63. Li Y, Fang Y, Fang J. Predicting residue-residue contacts using random forest models. *Bioinformatics* 2011,**27**:3379-3384.
64. Wolda H. Similarity indices, sample size and diversity. *Oecologia* 1981,**50**:296-302.
65. Polikar R. Ensemble learning. In: *Ensemble Machine Learning*: Springer; 2012. pp. 1-34.
66. Krogh A, Sollich P. Statistical mechanics of ensemble learning. *Physical Review E* 1997,**55**:811.
67. Guyon I, Elisseeff A. An introduction to variable and feature selection. *The Journal of Machine Learning Research* 2003,**3**:1157-1182.
68. Sinisi SE, Polley EC, Petersen ML, Rhee SY, van der Laan MJ. Super learning: an application to the prediction of HIV-1 drug resistance. *Stat Appl Genet Mol Biol* 2007,**6**:Article7.
69. Gama J, Brazdil P. Cascade generalization. *Machine Learning* 2000,**41**:315-343.
70. Saha I, Zubek J, Klingstrom T, Forsberg S, Wikander J, Kierczak M, *et al.* Ensemble Learning Prediction of Protein-Protein Interactions using Proteins Functional Annotations. *Molecular BioSystems* 2014.
71. Yang J, Jang R, Zhang Y, Shen HB. High-accuracy prediction of transmembrane inter-helix contacts and application to GPCR 3D structure modeling. *Bioinformatics* 2013,**29**:2579-2587.
72. Skwark MJ, Abdel-Rehim A, Elofsson A. PconsC: combination of direct information methods and alignments improves contact prediction. *Bioinformatics* 2013,**29**:1815-1816.
73. Hakes L, Lovell SC, Oliver SG, Robertson DL. Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proc Natl Acad Sci U S A* 2007,**104**:7999-8004.
74. Dutheil JY. Detecting coevolving positions in a molecule: why and how to account for phylogeny. *Brief Bioinform* 2012,**13**:228-243.
75. Ha JH, Loh SN. Protein conformational switches: from nature to design. *Chemistry* 2012,**18**:7984-7999.
76. Fodor AA, Aldrich RW. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* 2004,**56**:211-221.

77. Morikawa Y, Zhang WH, Hockley DJ, Nermut MV, Jones IM. Detection of a trimeric human immunodeficiency virus type 1 Gag intermediate is dependent on sequences in the matrix protein, p17. *J Virol* 1998,**72**:7659-7663.
78. Kiernan RE, Ono A, Freed EO. Reversion of a human immunodeficiency virus type 1 matrix mutation affecting Gag membrane binding, endogenous reverse transcriptase activity, and virus infectivity. *J Virol* 1999,**73**:4728-4737.
79. Tedbury PR, Ablan SD, Freed EO. Global Rescue of Defects in HIV-1 Envelope Glycoprotein Incorporation: Implications for Matrix Structure. *PLoS Pathog* 2013,**9**:e1003739.
80. Pornillos O, Ganser-Pornillos BK, Yeager M. Atomic-level modelling of the HIV capsid. *Nature* 2011,**469**:424-427.
81. Pornillos O, Ganser-Pornillos BK, Kelly BN, Hua Y, Whitby FG, Stout CD, *et al.* X-ray structures of the hexameric building block of the HIV capsid. *Cell* 2009,**137**:1282-1292.
82. Byeon IJ, Meng X, Jung J, Zhao G, Yang R, Ahn J, *et al.* Structural convergence between Cryo-EM and NMR reveals intersubunit interactions critical for HIV-1 capsid function. *Cell* 2009,**139**:780-790.
83. Yufenyuy EL, Aiken C. The NTD-CTD intersubunit interface plays a critical role in assembly and stabilization of the HIV-1 capsid. *Retrovirology* 2013,**10**:29.
84. Liang C, Hu J, Russell RS, Roldan A, Kleiman L, Wainberg MA. Characterization of a putative α-helix across the capsid-SP1 boundary that is critical for the multimerization of human immunodeficiency virus type 1 Gag. *Journal of virology* 2002,**76**:11729-11737.
85. Liu Y, Eyal E, Bahar I. Analysis of correlated mutations in HIV-1 protease using spectral clustering. *Bioinformatics* 2008,**24**:1243-1250.
86. Haq O, Levy RM, Morozov AV, Andrec M. Pairwise and higher-order correlations among drug-resistance mutations in HIV-1 subtype B protease. *BMC Bioinformatics* 2009,**10 Suppl 8**:S10.
87. Prabu-Jeyabalan M, Nalivaika E, Schiffer CA. Substrate shape determines specificity of recognition for HIV-1 protease: analysis of crystal structures of six substrate complexes. *Structure* 2002,**10**:369-381.
88. Lee SK, Potempa M, Kolli M, Ozen A, Schiffer CA, Swanstrom R. Context surrounding processing sites is crucial in determining cleavage rate of a subset of processing sites in HIV-1 Gag and Gag-Pro-Pol polyprotein precursors by viral protease. *J Biol Chem* 2012,**287**:13279-13290.
89. Vercauteren J, Beheydt G, Prosperi M, Libin P, Imbrechts S, Camacho R, *et al.* Clinical evaluation of Rega 8: an updated genotypic interpretation system that significantly predicts HIV-therapy response. *PLoS One* 2013,**8**:e61436.
90. Watanabe SM, Chen MH, Khan M, Ehrlich L, Kemal KS, Weiser B, *et al.* The S40 residue in HIV-1 Gag p6 impacts local and distal budding determinants, revealing additional late domain activities. *Retrovirology* 2013,**10**:143.
91. Datta SA, Curtis JE, Ratcliff W, Clark PK, Crist RM, Lebowitz J, *et al.* Conformation of the HIV-1 Gag protein in solution. *J Mol Biol* 2007,**365**:812-824.
92. Yates PJ, Hazen R, St Clair M, Boone L, Tisdale M, Elston RC. In vitro development of resistance to human immunodeficiency virus protease inhibitor GW640385. *Antimicrob Agents Chemother* 2006,**50**:1092-1095.
93. Prado JG, Wrin T, Beauchaine J, Ruiz L, Petropoulos CJ, Frost SD, *et al.* Amprenavir-resistant HIV-1 exhibits lopinavir cross-resistance and reduced replication capacity. *AIDS* 2002,**16**:1009-1017.
94. Mo H, Parkin N, Stewart KD, Lu L, Dekhtyar T, Kempf DJ, *et al.* Identification and structural characterization of I84C and I84A mutations that are associated with high-level resistance to human immunodeficiency virus protease inhibitors and impair viral replication. *Antimicrob Agents Chemother* 2007,**51**:732-735.
95. Nijhuis M, Wensing AM, Bierman WF, de Jong D, Kagan R, Fun A, *et al.* Failure of treatment with first-line lopinavir boosted with ritonavir can be explained by novel resistance pathways with protease mutation 76V. *J Infect Dis* 2009,**200**:698-709.
96. Shibata J, Sugiura W, Ode H, Iwatani Y, Sato H, Tsang H, *et al.* Within-host co-evolution of Gag P453L and protease D30N/N88D demonstrates virological advantage in a highly protease inhibitor-exposed HIV-1 case. *Antiviral Res* 2011,**90**:33-41.
97. Aoki M, Venzon DJ, Koh Y, Aoki-Ogata H, Miyakawa T, Yoshimura K, *et al.* Non-cleavage site gag mutations in amprenavir-resistant human immunodeficiency virus type 1 (HIV-1) predispose HIV-1 to rapid acquisition of amprenavir resistance but delay development of resistance to other protease inhibitors. *J Virol* 2009,**83**:3059-3068.

98. Gatanaga H, Suzuki Y, Tsang H, Yoshimura K, Kavlick MF, Nagashima K, *et al.* Amino acid substitutions in Gag protein at non-cleavage sites are indispensable for the development of a high multitude of HIV-1 resistance against protease inhibitors. *J Biol Chem* 2002,**277**:5952-5961.
99. Dam E, Quercia R, Glass B, Descamps D, Launay O, Duval X, *et al.* Gag mutations strongly contribute to HIV-1 resistance to protease inhibitors in highly drug-experienced patients besides compensating for fitness loss. *PLoS Pathog* 2009,**5**:e1000345.
100. Ho SK, Coman RM, Bunger JC, Rose SL, O'Brien P, Munoz I, *et al.* Drug-associated changes in amino acid residues in Gag p2, p7(NC), and p6(Gag)/p6(Pol) in human immunodeficiency virus type 1 (HIV-1) display a dominant effect on replicative fitness and drug response. *Virology* 2008,**378**:272-281.
101. Brann TW, Dewar RL, Jiang MK, Shah A, Nagashima K, Metcalf JA, *et al.* Functional correlation between a novel amino acid insertion at codon 19 in the protease of human immunodeficiency virus type 1 and polymorphism in the p1/p6 Gag cleavage site in drug resistance and replication fitness. *J Virol* 2006,**80**:6136-6145.
102. Maguire MF, Guinea R, Griffin P, Macmanus S, Elston RC, Wolfram J, *et al.* Changes in human immunodeficiency virus type 1 Gag at positions L449 and P453 are linked to I50V protease mutants in vivo and cause reduction of sensitivity to amprenavir and improved viral fitness in vitro. *J Virol* 2002,**76**:7398-7406.
103. Kolli M, Stawiski E, Chappey C, Schiffer CA. Human immunodeficiency virus type 1 protease-correlated cleavage site mutations enhance inhibitor resistance. *J Virol* 2009,**83**:11027-11042.
104. Knops E, Kemper I, Schulter E, Pfister H, Kaiser R, Verheyen J. The evolution of protease mutation 76V is associated with protease mutation 46I and gag mutation 431V. *AIDS* 2010,**24**:779-781.
105. Bally F, Martinez R, Peters S, Sudre P, Telenti A. Polymorphism of HIV type 1 gag p7/p1 and p1/p6 cleavage sites: clinical significance and implications for resistance to protease inhibitors. *AIDS Res Hum Retroviruses* 2000,**16**:1209-1213.
106. Knops E, Brakier-Gingras L, Schulter E, Pfister H, Kaiser R, Verheyen J. Mutational patterns in the frameshift-regulating site of HIV-1 selected by protease inhibitors. *Med Microbiol Immunol* 2012,**201**:213-218.
107. Lastere S, Dalban C, Collin G, Descamps D, Girard PM, Clavel F, *et al.* Impact of insertions in the HIV-1 p6 PTAPP region on the virological response to amprenavir. *Antivir Ther* 2004,**9**:221-227.
108. Kaufmann GR, Suzuki K, Cunningham P, Mukaide M, Kondo M, Imai M, *et al.* Impact of HIV type 1 protease, reverse transcriptase, cleavage site, and p6 mutations on the virological response to quadruple therapy with saquinavir, ritonavir, and two nucleoside analogs. *AIDS Res Hum Retroviruses* 2001,**17**:487-497.
109. Lambert-Niclot S, Flandre P, Malet I, Canestri A, Soulie C, Tubiana R, *et al.* Impact of gag mutations on selection of darunavir resistance mutations in HIV-1 protease. *J Antimicrob Chemother* 2008,**62**:905-908.
110. Cote HC, Brumme ZL, Harrigan PR. Human immunodeficiency virus type 1 protease cleavage site mutations associated with protease inhibitor cross-resistance selected by indinavir, ritonavir, and/or saquinavir. *J Virol* 2001,**75**:589-594.
111. Larrouy L, Charpentier C, Landman R, Capitant C, Chazallon C, Yeni P, *et al.* Dynamics of gag-pol minority viral populations in naive HIV-1-infected patients failing protease inhibitor regimen. *AIDS* 2011,**25**:2143-2148.
112. Zhang YM, Imamichi H, Imamichi T, Lane HC, Falloon J, Vasudevachari MB, *et al.* Drug resistance during indinavir therapy is caused by mutations in the protease gene and in its Gag substrate cleavage sites. *J Virol* 1997,**71**:6662-6670.
113. Roquebert B, Malet I, Wirden M, Tubiana R, Valantin MA, Simon A, *et al.* Role of HIV-1 minority populations on resistance mutational pattern evolution and susceptibility to protease inhibitors. *AIDS* 2006,**20**:287-289.
114. Verheyen J, Litau E, Sing T, Daumer M, Balduin M, Oette M, *et al.* Compensatory mutations at the HIV cleavage sites p7/p1 and p1/p6-gag in therapy-naive and therapy-experienced patients. *Antivir Ther* 2006,**11**:879-887.
115. Kolli M, Lastere S, Schiffer CA. Co-evolution of nelfinavir-resistant HIV-1 protease and the p1-p6 substrate. *Virology* 2006,**347**:405-409.

116. Malet I, Roquebert B, Dalban C, Wirden M, Amellal B, Agher R, *et al.* Association of Gag cleavage sites to protease mutations and to virological response in HIV-1 treated patients. *J Infect* 2007,**54**:367-374.
117. Chang MW, Oliveira G, Yuan J, Okulicz JF, Levy S, Torbett BE. Rapid deep sequencing of patient-derived HIV with ion semiconductor technology. *J Virol Methods* 2013,**189**:232-234.
118. Rossi AH, Rocco CA, Mangano A, Sen L, Aulicino PC. Sequence variability in p6 gag protein and gag/pol coevolution in human immunodeficiency type 1 subtype F genomes. *AIDS Res Hum Retroviruses* 2013,**29**:1056-1060.
119. Demšar J. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research* 2006,**7**:1-30.

Supplementary text S1: A new ensemble coevolution system for detecting HIV-1 protein coevolution

Contents

1 Ensemble coevolution system

Introduction We designed an ensemble coevolution system (ECS) to provide robust predictions of coevolving residues. Ensemble learning systems which combine different prediction methods have shown high prediction performance in many studies (see review in [27, 28]). For instance, XCS was made to improve the self-adaptation of evolutionary algorithms [29]. LCSE enhances the rule-based classification through the combination of reinforcement learning, evolutionary computing and heuristic approaches [30]. GAssist improves the performance of the ordinal classification through the assembling of several rule-based models [31]. A cascade generalization framework which combines naïve Bayesian classifiers, linear discriminant classifiers and decision trees could improve the classification accuracy compared to individual classifiers [32]. An ensemble system has recently been developed to improve the prediction of protein-protein interactions using the attributes collected mainly from gene ontology annotations [33]. This ensemble system integrates four machine learning methods (support vector machine, random forest, decision tree and naïve Bayesian network) based on the majority voting strategy [33]. Recently, an ensemble method which combines PSICOV and machine learning classifiers can improve the prediction of transmembrane inter-helix contacts [34].

An ensemble learning system is usually built to combine a set of prediction models and is popular when the prediction variability between prediction models (classifiers) is high [28]. Ensemble learning is not needed if all models predict the same results [28]. Ensemble learning systems usually comprise of three parts: (1) data sampling/selection, (2) model prediction, (3) a combiner. The combiner plays a key role to determine the strategy of how different predictions from various methods are integrated. There are many popular ensemble strategies such as: majority voting (predictions supported by more than 50% of methods), weighted voting (predictions are weighted according to the importance of models) and Borda count (predictions consistently obtained by all the models) [27, 28].

Inspired by the principle of ensemble learning, we endeavored to build a software system which integrates known sequence-based methods for coevolution prediction. It turned out to be difficult for several reasons. Firstly, there is a lack of gold standard datasets for training the ensemble learning system. Secondly, sequence-based methods predict different scores for statistical couplings and mostly do not predict true negatives, which limits our choices on data sampling and ensemble strategies. Thirdly, some sequence-based methods require heavy computational time, a limit which may restrain a broad application of ensemble learning on large protein families. For the above reasons, the potential ensemble coevolution system should provide robust predictions while combining the number of sequence-based

methods as few as possible. Moreover, we considered the design of an ensemble coevolution system as an optimization problem, where the objective function was defined as a linear function (see Methods). Here, we describe the details of our heuristic algorithm which we designed to identify the combination of sequence-based methods which improves the prediction performance.

Algorithm 1 A heuristic algorithm for identifying the combination of sequence-based methods

Require: As inputs, a set of sequence-based methods $M = \{M_i | i = 1, ..., N\}$ and multiple sequence datasets $D = \{D_i | i = 1, ..., T\}$.

Ensure: As an output, the optimized method combination Ω^+.

$\Omega = \phi$; {Initiate the method set Ω as empty.}
$f(\Omega, D) = 0$; {Initiate the performance score of Ω as 0 given the datasets D.}
$\Delta_i = 0, i = 0, ..., N$; {Initiate the performance increase as 0 for each method.}
Step 1: Apply sequence-based methods and perform the linear transformation.
for $i = 1$ to N **do**
  for $j = 1$ to T **do**
    Obtain the coevolution scoring matrix $C(M_i, D_j)$;
    Linear transformation: $C^*(M_i, D_j) = \frac{C(M_i,D_j) - min(C(M_i,D_j))}{max(C(M_i,D_j)) - min(C(M_i,D_j))}$;
  end for
end for
Step 2: Optimization of method combination.
while $\Omega = \phi$ or $max(\Delta) > 0$ **do**
  for $i = 1$ to N **do**
    $\Omega^* = \Omega \cup M_i$ where $M_i \in M/\Omega$; {Add the unvisited method M_i to Ω^*.}
    for $j = 1$ to T **do**
      $C^*(\Omega^*, D_j) = \sum_{M_k \in \Omega^*} w_{k,j} \cdot C^*(M_k, D_j)$; {Integrate the coevolution predictions.}
    end for
    $\Delta_i = \frac{1}{T} \sum_{j=1}^{T} f(C(\Omega^*, D_j)) - f(\Omega, D)$; {Increase of average performance score.}
  end for
  if $max(\Delta) > 0$ **then**
    $M^* = arg_{M_i} max(\Delta)$; {A method with the highest increase of performance score.}
    $\Omega = \Omega \cup M^*$; {Update the best method subset.}
    $f(\Omega, D) = f(\Omega, D) + max(\Delta)$; {Update the best performance score.}
  end if
end while
Return Ω as Ω^+.

Methodology $M = \{M_i | i = 1, ..., N\}$ represents a set of sequence-based methods M_i and

$D = \{D_j | j = 1, ..., T\}$ represents sequence datasets, where N is the number of methods and T is the number of sequence datasets. Given a training dataset D_j, the sequence-based method M_i predicts a coevoltuion score for each statistical coupling. This coevolution score is normalized in the matrix with $L \times L$ elements, denoted as $C^*(M_i, D_j)$, where L is the length of residue positions in D_j. Every non-empty element in this matrix represents the coevolution score of the statistical coupling between the positions n and m. To contend with varied statistical measurements used by different methods, four normalization strategies have been previously proposed [13]. Given a continuous variable x as an input, the normalized variable y satisfies:
(1) Linear transformation: $y = \frac{x - min(x)}{max(x) - min(x)}$.
(2) Power transformation: $y = 10^{(max(x)-x)(max(x)-min(x))}$.
(3) Binary transformation: if $Rank(n, m) < \alpha$, then $y = 1$; otherwise $y = 0$. The cutoff α is the number of top-ranked couplings ($\alpha = L$ in our analysis).
(4) Log transformation: $y = b \times x^a$ where $V_{max} = log_{10}[max(x)]$, $V_{min} = log_{10}[min(x)]$, if $min(x), max(x) > 0$, then $a = K/(V_{max} - V_{min}), b = -K \cdot V_{max}/(V_{max} - V_{min})$, and $K = 5$ as default.

We compared these four strategies and chose linear transformation as our normalization strategy because linear transformation performed the best using HIV-1 datasets (data not shown). Based on the normalized coevolution scores, the coupling (n, m) in the entire matrix is ranked, denoted as $Rank(n, m)$. Given a statistical measurement f (e.g. AUC), the performance of the sequence-based method M_i is measured by $f(C^*(M_i, D_j))$. Suppose w_i and u_j denotes the weight of the method M_i and the weight of the D_j, respectively. The objective function for the ensemble learning is defined as:

$$F(\Omega, D) = \sum_{j=1}^{T} \frac{u_j}{|\Omega|} \sum_{M_i \in \Omega} w_i \times f(C^*_{n,m}(M_i, D_j))$$

An optimized combination of methods Ω^+ is identified when $\Omega^+ = \max_{\Omega \in M} F(\Omega, D)$. To simplify the learning procedure with a small number of training datasets, our study assumed that $u_j = 1$ and w_i equals to 1 or 0. We therefore designed a heuristic algorithm (**Algorithm 1**) that provides a suboptimal solution to identify Ω^+ by maximizing the performance scores. Specifically, we used the forward selection to improve the prediction performance because of the high computational complexity. The forward selection each time adds one method into the optimized method set if the added method increases the performance score. Our heuristic algorithm begins with the initiation of global variables (line 1-3). The coupling predictions of sequence-based methods M are performed given the sequence dataset D (line

5-10, $T = 7, N = 27$ in our analysis). Given each method with a sequence dataset, statistical couplings in the scoring matrix $C(M_i, D_j)$ are obtained according to method measurements (line 7). The scores are then linearly transformed between 0 and 1 (line 8). Thereafter, the forward selection adds one method into the method subset Ω at each loop (line 14). It also assembles the statistical coupling predictions for the AUC evaluation (line 16, see AUC definition in Methods). The increase of performance score Δ_i is calculated for each method M_i when added into the method subset Ω (line 18). In each round, one method that increases the highest performance score is added into the method subset Ω (line 20-24). The procedure ends when adding any method does not improve the best performance score (line 12). The method subset Ω is returned as the optimized method combination Ω^+ (line 26).

The forward selection algorithm can be easily adapted into the backward elimination, which requires the initiate parameter $\Omega = \{1, ..., N\}$ and line 14 in the algorithm should remove a single method instead of adding a method. In order to achieve a promising optimization, we implemented both forward selection and backward elimination approaches. In our experiments, we found that both strategies identified the method set Ω with four methods (NCPS, RCW, PhyCMAP, CMPro), suggesting a convergence of the heuristic search. Overall, our heuristic algorithm offers a fast computation to identify a method combination with improved prediction performance using a local optimization procedure.

Our parameter settings: Parameters of sequence-based methods were initialized according to individual methods (see parameter settings in the next section).

Software availability Our toolbox.

Table 1. Summary of 27 sequence-based methods integrated in ECS

Method	Methodology	Software availability	Year	Ref
ASC	Mutual information	Our toolbox	2011	[10]
APC	Mutual information	Our toolbox	2011	[10]
BN	Bayesian network	https://code.google.com/p/bright/	2007	[22]
CTMP	Markov model, phylogenetic tree	http://www.stat.sinica.edu.tw/chyeang/	2007	[24]
CoMap	Compensation coefficient, phylogenetic tree	http://gna.org/projects/comap	2007	[25]
Complementary	Complementary matrix, Pearson coefficient	Our toolbox	2006	[11]
CMPro	Neural network	http://scratch.proteomics.ics.uci.edu/	2012	[19]
DNcon	Deep network, Bolzmann machine	http://iris.rnet.missouri.edu/dncon/	2012	[18]
GREMLIN	Maximum entropy function	http://openseq.org/	2013	[26]
Interdependency	Entropy, mutual information	http://www.uhnresearch.ca/labs/tillier/depend2	2004	[5]
LogR	Bayesian network, APC	Author's generosity	2010	[13]
MI	Mutual information	Our toolbox	-	[1]
MIBP	MI, physicochemical property	http://www.biomedcentral.com/1471-2105/12/206	2011	[7]
Mutagenetic	Maximum likelihood mixed tree	http://mtreemix.bioinf.mpi-inf.mpg.de/	2005	[21]
NBZPX2	Normal binary	Toolbox in [6]	2012	[6]
NCPS	Mutual information, sequence similarity	Our toolbox	2009	[9]
NNcon	Neural network	http://casp.rnet.missouri.edu/nncon.html	2009	[17]
PCC	Mutual information, Pearson coefficient	Our toolbox	2010	[12]
PhyCMAP	Random forest, integer linear programming	http://raptorx.uchicago.edu/	2013	[20]
plmDCA	Maximum entropy function	http://plmdca.csc.kth.se/	2013	[35]
PSICOV	Sparse inverse covariance	http://bioinfadmin.cs.ucl.ac.uk/downloads/PSICOV/	2012	[15]
PhysicoMI	MI, AA substitution matrix	Our toolbox	2012	[8]
RCW	Mutual information	Our toolbox	2007	[3]
SCA	Statistical free energy coupling	http://systems.swmed.edu/rr_lab/sca.html	2009	[10]
Spidermonkey	MCMC Bayesian network, phylogenetic tree	http://www.hyphy.org/w/index.php/Main_Page	2008	[23]
SVMcon	Support vector machine	http://casp.rnet.missouri.edu/svmcon.html	2006	[16]
ZRES	Mutual information	Toolbox in [6]	2009	[4]

2 Position-specific sequence-based methods in the last decade

This section provides more details about the 27 sequence-based methods (Table 1). For each method, we begin with a simple introduction and then briefly describe their key mathematical models. Lastly, we explain the parameter settings used in this study and the software availability. We order these methods according to their methodology so that methods with similar methodologies are described together.

2.1 MI: mutual information [1]

Introduction Mutual information (MI) measures the contribution of the knowledge of variable X's information in the redution of the uncertainty of the other variable Y [36]. For its simplicity, MI has been adapted to predict coevolving positions and protein contact map [37].

It was proposed based on the hypothesis that coevolving residues or residues in contact tend to share a high mutual information [1].

Methodology Let variables X, Y represent two protein positions in multiple sequence alignment (MSA) from a protein (family), $X = x$ indicates that the position X takes the amino acid x which is one of amino acid forms in MSA, $P(X = x)$ represents the marginal probability of the position X taking the amino acid form x in MSA. Likewise, $P(X = x, Y = y)$ is the joint probability that the position X takes amino acid form x and Y takes amino acid form y simultaneously. The mutual information between $X = x$ and $Y = y$ is defined as:

$$MI(X = x, Y = y) = P(X = x, Y = y) \log \frac{P(X = x, Y = y)}{P(X = x)P(Y = y)} \tag{1}$$

Furthermore, the mutual information between two positions X, Y is the sum of mutual information of all possible configurations at the position X and Y, defined as:

$$MI(X, Y) = \sum_x \sum_y P(X = x, Y = y) \log \frac{P(X = x, Y = y)}{P(X = x)P(Y = y)} \tag{2}$$

Our parameter settings In cases where the marginal probability $P(X = x)$ or $P(Y = y)$ in the denominator was zero causing the impossible infinity in the log function of MI, Laplace smoothing was used by adding 1 into the both denominator and the numerator of the marginal probability $P(X)$ ($P(X = x) = n/N \rightarrow P(X = x) = (n + 1)/(1 + N)$) [38]. Gaps from the positions of interest were ignored in the calculation of probabilities [1]. To solve the heavy computation in large protein families, a fast parallel programming code is provided in our Toolbox.

Software availability Our toolbox and toolbox in [6].

2.2 ASC and APC: average sum and product correction [2] (2008)

Introduction In this study, two statistical methods were proposed to estimate the background MI in the process of protein evolution. The background MI was defined as the average MI over all position pairs [2]. The average sum correction (ASC) was defined as the sum of MI calculated at the positions i and j minus the background MI [2]. Similarly, the average product correction (APC) was defined as the product of MI calculated calculated at the positions i and j divided by the background MI.

Methodology Based on mutual information described in the previous section, the background mutual information in ASC [2] is defined as:

$$ASC_{Backgroud}(i,j) = \frac{1}{2n}\sum_{X}[MI(i,X) + MI(X,j)] - \frac{1}{n^2}\sum_{X}\sum_{Y\neq X} MI(X,Y) \quad (3)$$

Suppose i and j are two residue positions of interest and n is the number of the overall protein positions. Given an input MSA, the second part of above formula is a constant so that the ASC correction for mutual information is defined as:

$$ASC_{MI}(i,j) = MI(i,j) - \frac{1}{2n}\sum_{X}[MI(i,X) + MI(X,j)] \quad (4)$$

Assuming that the background dependency is a product of independent factors associated with two positions, the average product correction (APC) is defined as:

$$APC_{MI}(i,j) = MI(i,j) - \frac{\sum_{X} MI(i,X) \times \sum_{X} MI(X,j)}{\sum_{X,Y} MI(X,Y)} \quad (5)$$

Parameter settings We used the same parameter settings as the mutual information.

Software availability Our toolbox.

2.3 RCW: row and column weighted MI [3] (2007)

Introduction Similar to the methodology of ASC and APC, the method RCW takes the average mutual information as the "weight" for the pairwise dependency. Given simulated datasets, RCW outperformed MI, logarithm correlation and multi-dimensional amino acid representation [3].

Methodology Let i and j be the positions of interest, $MI(i,j)$ denotes the mutual information between the positions i and j. The mean value of mutual information at the position i is calculated as $\overline{MI_i} = \sum_{j=1}^{n} MI(i,j)/(n-1)$, then RCW between the position pair (i,j) is defined as:

$$RCW(i,j) = \frac{MI(i,j)}{\overline{MI_i} + \overline{MI_j} - 2MI(i,j)/(n-1)} \quad (6)$$

Our parameter settings We used the same parameter settings as those in mutual informa-

tion.

Software availability `http://www.cbs.dtu.dk/services/InterMap3D/`.

2.4 ZRES: Z-residue score [4] (2009)

Introduction This study provided a method which refines MI by removing strong non-coevolutionary influence and accounting for the position variability. The method is built based on the linear regression between the mutual information MI_{ij} and $\overline{MI_i} \times \overline{MI_j}$. Using protein sequences from 1592 protein families in the Pfam database (http://pfam.sanger.ac.uk/), this study showed that predicted coevolving positions tend to be in a close physical proximity [4].

Methodology As the stochastic and phylogenetic bias may affect the performance of MI, ZRES uses the linear regression to fit the mutual information MI_{ij} with $\overline{MI_i} \times \overline{MI_j}$. By doing so, the biases can be measured by $Res_{ij} = \overline{MI_i} \times \overline{MI_j} - \beta \cdot MI_{ij}$, where β is the estimated coefficient in the linear regression. Based on this principle, the statistical coupling between the positions i and j is quantified by the Z-score, denoted as $ZRes(i, j)$:

$$ZRes(i,j) = \frac{(Res_{ij} - \frac{1}{n}\sum_{j=1}^{n} Res_{ij})(Res_{ij} - \frac{1}{n}\sum_{i=1}^{n} Res_{ij})}{\sqrt{\sum_{i=1}^{n}(Res_{ij} - \frac{1}{n}\sum_{j=1}^{n} Res_{ij})^2}\sqrt{\sum_{j=1}^{n}(Res_{ij} - \frac{1}{n}\sum_{i=1}^{n} Res_{ij})^2}} \tag{7}$$

The higher the Z-score $ZRes(i, j)$, the higher the chance that two positions i and j are coevolving.

Our parameter settings We used the default parameter settings in the ZRES toolbox [6].

Software availability Toolbox in [6].

2.5 Interdependency V2.1 [5] (2004)

Introduction To reduce phylogenetic bias, the statistical couplings in this method were quantified by the statistical interdependency ratio, which measured the differences between the observed residue interdependency and the expected residue interdependency [5].

Methodology The expected independency is estimated by the likelihood of equivalent residues compared to all residues at other positions. Let N be the number of non-gap residues at the position X_i, the expected interdependency of the position X_i is defined as:

$$MS(X_i) = \frac{1}{N} \sum_{j \neq i} \sum_{x_i, x_j} \log \frac{P(X_i = x_i, X_j = x_j)}{P(X_i = x_i) P(X_j = x_j)} \tag{8}$$

The interdependency is proposed using the entropy weight based on the hypothesis that residue positions do not increase the dependency when no functional correlation exists [5]. To quantify the amino acid variation at the position X_i, the entropy function $H(X_i)$ is measured through $H(X_i) = -\sum_{x_i} P(X_i = x_i) \log P(X_i = x_i)$. Based on the mutual informaiton and the entropy function, the interdependency ratio is defined as:

$$R(X_i, X_j) = \frac{MI(X_i, X_j) H(X_i) H(X_j)}{MS(X_i) + MS(X_j)} [1 - H(X_i) H(X_j)] \tag{9}$$

Our parameter settings We performed the analyses using the default statistical parameters.
Software availability `http://www.uhnresearch.ca/labs/tillier/depend2/dependency.html`

2.6 NBZPX2: normal binary ZPX2 [6] (2012)

Introduction The method NBZPX2 which is an adaption of ZRES improves the performance of ZRES by refining the MSA inputs. The refinement strategy uses the data transformation called the normal binary [6].

Methodology The workflow of this method can be simply described by three steps. (1) The sequences in the MSA inputs are reordered using the sequence similarity, i.e. the 1st and 2nd sequences are the most similar sequences and the 3rd sequence is the the closest one to the 2nd, etc. This reordering process terminates until all sequences have been visited. (2) The MSA input is transformed into a binary dataset through AA comparisons of two subsequent sequences. Specifically, the binary value 1 indicates that two AAs are identical, otherwise 0. (3) The ZRES algorithm is applied using the transformed dataset.

Our parameter settings We used the default parameter settings.

Software availability Toolbox in [6].

2.7 MIBP: MI with physicochemical property [7] (2011)

Introduction This study proposed a covariation model which uses mutual information accounting for residue physicochemical properties [7]. Ten groups of amino acids were classified according to physicochemical properties. The key concept of this model is to calculate the mutual information between amino acid groups, while taking into account the background amino acid distribution. Performance of this model was compared to MI and ELSC using three protein families (1JXA-A, 1B93-A and PF01053).

Methodology Based on MI, the MIBP covariation between the positions i and j is defined as:

$$MIBP(i,j) = \sum_{a_n} \sum_{b_m} P(x_i \in a_n, x_j \in b_m) \log \frac{P(x_i \in a_n, x_j \in b_m)}{P_b(x_i \in a_n) P_b(x_j \in b_m)}$$

Where x_i and x_j are residues at the positions i and j in the MSA input, a_n (n=1...10) and b_m (m=1...10) denotes the AA functional group at the positions i and j, respectively. Ten residue groups include hydrophobic (A, G, C, T, I, V, L, K, H, F, Y, W, M), aromatic (F, Y, W, H), aliphatic (I, V, L), tiny (A, S, G, C), small (P, N, D, T, C, A, G, S, V), proline (P), charged (K, H, R, D, E), negative (D, E), polar (N, Q, S, D, E, C, T, K, R, H, Y, W) and positive (K, H, R). Based on BLOSUM62 substitution matrix, $P_b(x_i \in a_n)$ is the background distribution of physicochemical properties and is defined as:

$$P_b(x_i \in a_n) = \frac{P(x_i \in a_n)/B(a_n)}{\sum_{a_n} P(x_i \in a_n)/B(a_n)}$$

Where $B(a_n)$ denotes the BLOSUM62 constants for functional groups (hydrophobic: 0.504, aromatic: 0.132, aliphatic: 0.111, tiny: 0.243, small: 0.6632, charged: 0.226, negative: 0.117, positive: 0.507, proline: 0.244, polar: 0.043) [7].

Our parameter settings We used the default parameters in the original python implementation.

Software availability `http://www.biomedcentral.com/1471-2105/12/206`.

2.8 PhysicoMI: physico-chemical corrected MI [8] (2012)

Introduction PhysicoMI was proposed to calculate residue similarities taking into account physical-chemical properties. In this method, MI and AA frequency are corrected using the

BLOSUM62 substitution matrix [8].

Methodology Suppose x, y are two AAs at the positions X and Y respectively, the corrected joint probability of X in the presence of Y is modeled as:

$$f(X = x, Y = y) = \frac{n(x, y) + \sum_{(x',y') \neq (x,y)} n(x', y') S(x, x') S(y, y') / \sqrt{N}}{N + \sqrt{N}} \tag{10}$$

Where $n(x, y)$ is the number of residues x at the position X and y at the position Y. $S(x, x')$ is the substitution score when amino acid x' is replaced by x. In the same fashion, the corrected marginal probability is defined as:

$$f(X = x) = \frac{n(x) + \sum_{x' \neq x} n(x') S(x, x') / \sqrt{N}}{N + \sqrt{N}} \tag{11}$$

In the final measurement, both marginal and joint probabilities are corrected by physical-chemical properties:

$$MI_{Physico}(X, Y) = \sum_{X=x} \sum_{Y=y} f(x, y) \log \frac{f(X = x, Y = y)}{f(X = x) f(Y = y)} \tag{12}$$

Our parameter settings We used the default parameter settings.

Software availability Our toolbox.

2.9 NCPS: normalized coevolutionary pattern similarity [9] (2009)

Introduction To identify coevolving positions by MI can be complicated due to common ancestry and stochastic noise [9]. NCPS was therefore proposed to normalize sequence similarities by reducing the background noise in the correlated mutation analysis. This study showed that the background noise could be reduced using three coevolution analyses: McBASC, OMES and MI [9].

Methodology Suppose $CM(i, j)$ represents the correlated mutation score between the positions i and j. The coevolutionary pattern similarity (CPS) between the positions i and j is modeled by the dot product of two vectors. Let n be the number of overall residue positions

in the sequences, $CPS(i,j)$ is defined as:

$$CPS(i,j) = \frac{1}{n-2} \sum_{k \neq i,j} CM(i,k)CM(j,k) \tag{13}$$

Secondly, the coevolutionary pattern similarity is normalized as follows:

$$NCPS(i,j) = \frac{CPS(i,j)}{\sqrt{\frac{1}{n(n-1)} \sum_{i,j} CPS(i,j)}} \tag{14}$$

Thirdly, the NCPS score is proposed by removing the background noise.

$$aMIc(i,j) = \frac{1}{2}\left[\frac{MI(i,j) - NCPS(i,j)}{\max[MI(i,j) - NCPS(i,j)]} + \frac{E(i,j)[MI(i,j) - NCPS(i,j)]}{\max\left[E(i,j)[MI(i,j) - NCPS(i,j)]\right]}\right] \tag{15}$$

Where $E(i,j) = H(i)H(j)[1 - H(i)H(j)]$ is the entropic factor [9].

Our parameter settings We used the default parameter settings.

Software availability Our toolbox.

2.10 SCA: statistical coupling analysis [10] (2009)

Introduction Statistical coupling analysis (SCA) has been shown to reveal allosteric communications by the energetically coupled positions in the PDZ protein family [39]. Later studies showed that SCA could discover evolutionary networks which mediate the allosteric communications [40]. Moreover, SCA can be useful for protein design. For instance, artificial WW domains were designed based on the statistical couplings predicted by SCA. The artificial WW domains can be folded and bind with peptides in high affinities as natural WW domains [41].

Several follow-up studies have been devoted to improve the performance of SCA using different protein families [10,42,43]. The SCA toolbox written in Matlab® has been updated to Version 5.0, including different functions such as coupling prediction, independent component analysis and spectral decomposition. The latest version of SCA is available online.

Methodology SCA measures the statistical couplings observed in the functional interactions

of protein families [39]. The hypothesis relies on the observation that the distribution of amino acids in one position shifts due to the changes of amino acid distribution at another position. The degree of evolutionary dependence is quantified by the statistical coupling energy based on the Boltzmann equation [39]. The high coupling energy corresponds to the increased dependence between coevolving residues. Specifically, the statistical coupling energy between the positions i and j, denoted as $\Delta\Delta G_{i,j}$, is modeled as:

$$\Delta\Delta G_{i,j} = \sqrt{\sum_{x} (\ln P^{x}_{i|\delta j} - \ln P^{x}_{i})^2} \tag{16}$$

Where P^{x}_{i} is the probability of the residue x at the position i; $P^{x}_{i|\delta j}$ is the probability of the residue x at the position i given the perturbation position δj. The method in SCA V5.0 extends the coevolution estimation by the covariance analysis and principle component analysis [42]. Briefly, the pairwise correlation $C^{(ab)}_{ij}$ between the residue a at the position i and the residue b at the position j is modeled as:

$$C^{(ab)}_{ij} = \ln \Big[\frac{f^{(a)}_{i}(1-q^{(a)})}{(1-f^{(a)}_{i})q^{(a)}}\Big](f^{(ab)}_{ij} - f^{(a)}_{i} f^{(b)}_{j}) \ln \Big[\frac{f^{(b)}_{j}(1-q^{(b)})}{(1-f^{(b)}_{j})q^{(b)}}\Big] \tag{17}$$

Where $f^{(a)}_{i}$ is the frequency of having the residue a at the position i, $f^{(ab)}_{ij}$ is the joint frequency of having the residue a at the position i and the residue b at the position j, $q^{(a)}$ is the background probability of residue a in all proteins and q = (0.073, 0.025, 0.050, 0.061, 0.042, 0.072, 0.023, 0.053, 0.064, 0.089, 0.023, 0.043, 0.052, 0.040, 0.052, 0.073, 0.056, 0.063, 0.013, 0.033) (alphabetic order for the 20 amino acids).

Our parameter settings We used default parameter settings in the SCA toolbox V5.0 [42].

Software availability `http://systems.swmed.edu/rr_lab/sca.html`.

2.11 Complementary: complementary matrix in Pearson coefficient [11] (2006)

Introduction To predict inter-protein residue coevolution, this study proposed a method which calculates Pearson's coefficients accounting for the complementary residues between protein interaction interfaces [11]. It is known that residue frequencies and residue pairs at protein-protein interfaces follow certain complementary patterns [44]. For instance, abun-

dant hydrophobic residue pairs are often found at large protein interaction interfaces, while polar residue pairs usually occur at small interfaces [44]. Integrated into Pearson's coefficients, the complementary information was proven useful for inter-protein coevolution predictions. A promising performance of this method was found in the comparison with four other methods (MI, SCA, ELSC and OMES) using a sequence dataset containing 224 protein families in the Pfam database.

Methodology Let N be the number of sequences in the MSA input, S_i be the AA exchange matrix at the position i, $S_i(k,l)$ be the exchange score between the k^{th} and l^{th} residues at position i, $\overline{S_i}$ and σ_i be the mean and the standard deviation of residues in the exchange matrix S_i at the position i, respectively. Given the k^{th} sequence, $C_{i(k),j(k)}$ is the estimated complementary value between the k^{th} residues at the position i and j [44]. The corrected Pearson's coefficient between the position i and j, termed $r_{i,j}$, is modeled as:

$$r_{i,j} = \frac{1}{N^2} \sum_{k,l=1}^{N} \frac{(S_{i(k,l)} - \overline{S_i})(S_{j(k,l)} - \overline{S_j}) \times C_{i(k),j(k)} \times C_{i(l),j(l)}}{\sigma_i \sigma_j} \tag{18}$$

Our parameter settings The complementary matrix of 20 amino acids in [44] was used and the other parameters were default.

Software availability Our toolbox.

2.12 PCC: Pearson correlation coefficient [12] (2010)

Introduction To reduce background noise and phylogenetic bias, this study proposed Pearson's correlation coefficients (PCC) for statistical coupling predictions [12].

Methodology Suppose N is the number of sequences in the MSA input, S_i is the exchange matrix at the position i, $S_i(k,l)$ is the AA exchange score between the k^{th} and the l^{th} residues at the position i, $\overline{S_i}$ and σ_i are the mean and standard deviation of residues in the exchange matrix S_i, respectively. The Pearson's coefficient between the positions i and j, termed $r_{i,j}$,

is proposed as:

$$r(i,j) = \frac{1}{N^2} \sum_{k,l=1}^{N} \frac{(S_{i(k,l)} - \overline{S_i})(S_{j(k,l)} - \overline{S_j})}{\sigma_i \sigma_j} \tag{19}$$

$$\overline{S_i} = \frac{1}{N^2} \sum_{k,l=1}^{N} S_{i(k,l)}, \qquad \sigma_i = \sqrt{\frac{1}{N^2 - 1} \sum_{k,l=1}^{N} [S_{i(k,l)} - \overline{S_i}]^2} \tag{20}$$

Using the ASC to reduce the phylogenetic bias, the significance of statistical coupling is modeled as:

$$PCC(i,j) = r(i,\overline{x}) + r(j,\overline{x}) - \overline{r} \tag{21}$$

$$r(i,\overline{x}) = \frac{1}{N} \sum_{j=1}^{N} r(i,j), \qquad \overline{r} = \frac{1}{N} \sum_{i=1}^{N} r(i,\overline{x}) \tag{22}$$

Our parameter settings We used the default parameter settings.

Software availability Our toolbox.

2.13 LogR: disentangling direct coupling analysis [13] (2010)

Introduction The sequence-based method LogR was proposed to model the weighted covariations by disentangling indirect statistical dependencies from direct dependencies. Specifically, it quantifies the statistical couplings by estimating the weights of pairwise edges in Bayesian spanning trees, which can model the dependencies between residue positions [13]. In this study, the coevolving residue chains were found to travel through spatial distances in protein 3D structures, indicating the indirect (or transitive) statistical dependencies. To reduce phylogenetic biases, LogR used the phylogenetic correction proposed by APC . Moreover, the statistical dependency was estimated using informative prior and conservation information, both of which could improve the accuracy of contact predictions [13].

Methodology Suppose the number of residue positions is N given the MSA input D. Measured through Dirichlet prior, $P(D_{i,j})$ is the joint probability of the position i and j. The statistical dependency between the i^{th} and j^{th} positions, termed as $R_{i,j}$, is defined by the

joint probability $P(D_{i,j})$ divided by the marginal probability $P(D_i)$ and $P(D_j)$.

$$R_{i,j} = \frac{P(D_{i,j})}{P(D_i)P(D_j)} \tag{23}$$

To avoid the existence of 0 in the statistical independency, $\log R_{i,j}$ is shifted to a non-negative value by dividing the minimal value of $\log R_{i,j}$, which is $S_{i,j} = \log(R_{i,j}/\min R_{i,j})$. The APC phylogenetic correction is then calculated as:

$$\log(R^C_{i,j}) = S_{i,j} - \frac{\sum_{n=1}^{N} S_{n,i} \sum_{m=1}^{N} S_{m,j}}{\sum_{n=1}^{N}\sum_{m=1}^{N} S_{n,m}} \tag{24}$$

To disentangle the statistical couplings, the weight of the edge $j \leftarrow \pi(j)$ in the entire spanning tree space is estimated using the priors $P(\pi)$, where $\pi(j)$ is the parent node of the j^{th} position in the spanning tree π. The weighted correlation for the positions j and $\pi(j)$ is measured as:

$$M_{j,\pi(j)} = (R^C_{j,\pi(j)})^{\alpha} \frac{\mu_{j,\pi(j)}}{1 - \mu_{j,\pi(j)}} \tag{25}$$

Where $\pi(j)$ is the neighboring node of the node j^{th} in the spanning trees and $\mu_{j,\pi(j)}$ is the probability of the edge $j - \pi(j)$ in random spanning trees.

Our parameter settings We used the default parameter settings in the software.

Software availability Author's generosity.

2.14 DCA: direct coupling analysis [14] (2011)

Introduction This study aimed at the prediction of residue couplings in the spatial proximity given folded proteins [14]. Specifically, it approximates the maximum entropy by exploring the pairwise couplings given a MSA input. The implementation in [14] was proven to be fast compared to a message-passing algorithm published in an early study [45]. Moreover, the true positive rates of the new DCA has been shown to be better than MI and LogR using 131 domain families collected from the Pfam database.

Methodology Suppose $P(X_1, X_2, ..., X_n)$ is the join distribution given the MSA input with

the residue positions X_i from 1 to n. Given a maximum-entropy function, the optimization of the joint distribution is approximated by the marginal and the pairwise dependencies using the Lagrange transformation. Specifically, the approximation strategies (i.e. independent positions, mean-field approximation) are used to determine parameters in the Gibbs potential functions. Similar to mutual information, DCA models the pairwise couplings using the direct information (DI):

$$DI_{ij} = \sum_{x_i} \sum_{x_j} P_{ij}^{(dir)}(x_i, x_j) \cdot \log \left[\frac{P_{ij}^{(dir)}(x_i, x_j)}{\sum_{x_i} P_{ij}^{(dir)}(x_i, x_j) \cdot \sum_{x_j} P_{ij}^{(dir)}(x_i, x_j)} \right] \quad (26)$$

Where x_i is the residue at the position i, $P_{ij}^{(dir)}(x_i, x_j)$ is estimated through the following Gibbs potential function:

$$P_{ij}^{(dir)}(x_i, x_j) = \frac{1}{Z_{ij}} exp[-(f_{ij}(x_i, x_j) - f_i(x_i) f_j(x_j))^{-1}(x_i, x_j) + \widetilde{h}_i(x_i) + \widetilde{h}_j(x_j)] \quad (27)$$

Where f_i is the marginal probability function of the position i^{th} in the MSA input and f_{ij} is the joint distribution between the positions i and j. $\widetilde{h}_i(x_i)$ is the parameter that imposes the empirical single-residue counts of residue x_i at the position i. Z_{ij} is the normalization parameter. $(f_{ij} - f_i f_j)^{-1}(x_i, x_j)$ is the element in the inverse of an empirical correlation matrix derived from the MSA input.

Our parameter settings We used the default parameter settings.

Software availability `http://plmdca.csc.kth.se/`.

2.15 PSICOV: precise structural contact prediction [15] (2012)

Introduction PSICOV uses a graphical Lasso approach with a sparse inverse covariance estimation to reduce prediction biases, caused by functionally related residue chains in protein structures [46]. The rationale of the sparse inverse covariance method relies on the fact that residue contacts are sparse in the known protein structures. Specifically, the non-zero terms in the sparse inverse covariance matrix represent coupling positions and the zero terms indicate that two positions are conditionally independent, assuming that the underlying distribution follows a multivariate Gaussian distribution [46].

Methodology The methodology of PSICOV can be simply described as:

$$\Theta = [\lambda \cdot diag(\overline{X_1}, ..., \overline{X_N}) + (1 - \lambda) \cdot Cov(X, Y)]^{-1} \tag{28}$$

Where N is the number of residues in a sequence, $Cov(X, Y)$ is the covariance matrix over all positions X, Y in the MSA input, Θ is the concentration matrix, $\lambda \in [0, 1]$ is the shrinkage parameter which targets diagonal values in $diag(\overline{X_1}, ..., \overline{X_N})$ and $\overline{X_i}$ denotes the mean of diagonal values in the covariance matrix. The inverse covariance matrix estimates the significance of the positions i and j in contact through the function $S_{ij}^C = \sum_{ab} |\Theta_{ij}^{ab}|$, where a and b are two residues at the position i and j, respectively. The correction of phylogenetic bias has also been taken into account in the PSICOV score:

$$PC_{ij} = S_{ij}^C - \frac{\sum_{i=1}^{N} S_{ij}^C \cdot \sum_{j=1}^{N} S_{ij}^C}{\sum_{i=1}^{N} \sum_{j=1}^{N} S_{ij}^C} \tag{29}$$

Our parameter settings Recommended by the software manual, the parameter $-r$ and $-i$ were set to 0.005 and 62, respectively. Other parameters were default.

Software availability `http://bioinfadmin.cs.ucl.ac.uk/downloads/PSICOV/`.

2.16 SVMcon: support vector machine contact map predictor [16] (2007)

Introduction SVMcon was designed to predict residue-residue contacts using support vector machines (SVM). It was ranked as the second best method in the 7th critical assessment of structure prediction (CASP7). SVMcon also outperformed CMAPpro using the benchmark datasets.

Methodology SVMcon begins with the assessment of five input features for each residue pair at the positions i and j. Using over 310000 training data points, the input features include the local window features, pairwise information features, residue type features, central segment window features and protein information features. Thereafter, the classification function of

the SVM learner $f(x)$ is defined for contact predictions:

$$f(x) = \sum_{x_i \in S^+} \alpha_i \cdot K(x, x_i) - \sum_{x_i \in S^-} \alpha_i \cdot K(x, x_i) + b \tag{30}$$

$$K(x, x_i) = e^{-0.0025 \cdot ||x - x_i||^2} \tag{31}$$

Where x_i is the data point in a support vector, b represents the estimation bias, α_i is a non-negative weight assigned to the training data point x_i by minimizing a quadratic objective function. S^+ indicates the data point of residue pairs in contact and S^- indicates that the data point of a residue pair which is not in contact. A new data point x is predicted to be positive if $f(x) > 0$. $K(x, x_i)$ is the Gaussian radial basis kernel and the inverse of the variance parameter is optimized to be 0.0025.

Our parameter settings We used the default parameter settings.

Software availability `http://casp.rnet.missouri.edu/svmcon.html`

2.17 NNcon: neural network-based contact map predictor [17] (2009)

Introduction The 2D-Recursive Neural Network (2D-RNN) models were trained to predict residue-residue contacts using a sequence dataset consisted of 482 protein families.

Methodology Ten 2D-RNNs are trained and validated using the 10-fold cross-validation on a sequence dataset with 482 proteins. These 10 models are assembled to predict residues in contact. The residue contacts are those residue pairs which have the Euclidean distance of their C_α atoms less than 8 or 12 angstroms. Given a protein sequence with the length of n amino acids, the 2D-RNN constructs a $n \times n$ input matrix and outputs a probability matrix with $n \times n$ contact elements.

Our parameter settings We used the default parameter settings.

Software availability `http://casp.rnet.missouri.edu/nncon.html`

2.18 DNcon: neural network contact prediction [18] (2012)

Introduction DNcon was proposed to improve the predictions of residue-residue contacts based on deep networks and boosting techniques [18]. Using the standard back propagation

algorithm, the weights of multiple layers in deep networks were trained based on the predicted secondary structures, solvent accessibility, amino acid features and position-specific scoring matrix [18]. The method was evaluated using the D329 dataset containing 329 proteins, the SVMCON_TEST dataset containing 48 proteins and the CASP9 dataset containing 16 proteins. The comparison experiments showed that DNcon outperformed two other methods, ProC_S3 and SVMcon.

Methodology Amino acid features (e.g. electrostatic charge, codon diversity, volume, polarity, secondary structure) are modeled in deep network classifiers with the combination of restricted Bolzmann machines. The short, medium and long range residue-residue pairs are sampled from a large database, which is used to train deep network classifiers with the improved prediction power. The final contact predictions are evaluated by the scoring function.

Our parameter settings We used the software server with the option of the top $5L$ predictions where L is the length of amino acids given an input sequence.

Software availability `http://iris.rnet.missouri.edu/dncon/`.

2.19 CMPro: 2D recursive neural network [19] (2012)

Introduction This study proposed a contact prediction architecture based on neural networks and structural alignment models. Using both CASP8 and CASP9 datasets, performance of CMPro was shown to outperform PSICOV and other methods which were tested in CASP8 and CASP9 contact prediction [19].

Methodology Neural network prediction models are constructed in three steps. Firstly, the coarse contacts and the orientations between secondary structure elements are predicted using 2D recursive neural networks. The probability of parallel contact, anti-parallel contact or no-contact is estimated using feature variables extracted from protein secondary structures and amino acid compositions in the sequence input. Secondly, the energy-based method is used to optimize the amino acid alignment of strand - strand and helix-helix secondary structures. The log-likelihood objective function is defined as:

$$E_A = -\sum_{i=1}^{n} \log P_A(\hat{a}_i, \hat{\theta}_i) \tag{32}$$

Where n is the number of anti - parallel (or parallel) contacting residue pairs, $\hat{a}_i$ and $\hat{\theta}_i$ are the true shift and phase for the ith example, respectively. Thirdly, a deep neural network architecture refines the prediction of residue contacts. Deep NN architecture has k layers and each layer contains $25 \times 9 \times 2$ residue features, $3 \times 7 \times 7$ coarse features, $4 \times 7 \times 7$ alignment features and 81 temporal features. Due to the heavy computation of backpropagated gradients in multi-layered neural networks, an incremental approach has been proposed to train the weights of neural networks. The prediction performance is further improved using 10-fold cross-validation.

Our parameter settings We used the default parameter settings.

Software availability `http://scratch.proteomics.ics.uci.edu/`

2.20 PhyCMAP: random forest, integer linear programming [20] (2013)

Introduction This study proposed a sequence-based method PhyCMAP, which integrates both evolutionary and physical restraints using random forests and integer linear programming approach [20]. The performance comparison showed that PhyCMAP outperformed NNcon, CMAPpro and DCA given the CASP10 dataset.

Methodology PhyCMAP has two components. The first component predicts the probability score of the residue contacts using random forests. The contact score for the position pair (i, j) is defined as:

$$HPS(i,j) = \sum_{h \in H} PS_{beta}(a_i^h, a_j^h) + PS_{helix}(a_i^h, a_j^h) \tag{33}$$

Where a_i^h is the residue in a homology sequence h aligned to the residue i in the query sequence, $PS_{beta}(a_i^h, a_j^h)$ is the probability of a residue pair (a_i^h, a_j^h) forming a contact in the β-sheet structure, $PS_{helix}(a_i^h, a_j^h)$ is the probability of residues a_i^h, a_j^h forming a contact connecting two helix structures. Both $PS_{helix}(a_i^h, a_j^h)$ and $PS_{beta}(a_i^h, a_j^h)$ are obtained from protein structures in the training dataset containing 900 non-redundant protein structures

The second component selects a set of top-ranked contacts by using the integer linear programming, which maximizes accumulative probabilities under a set of physical constraints.

$$\max_{X,R} \sum_{6 \leq j-i} (X_{i,j} \times HSP(i,j)) - g(R) \tag{34}$$

Where $X_{i,j}$ is a binary variable and $X_{i,j} = 1$ if the position pair (i, j) are in contact. $g(R) = \sum R_r$ is a linear penalty function with the parameter r defined over 8 hard and soft constrains. These constraints are mainly proposed based on the observations of residue contacts between two β-strand structures or between two α-helix structures.

Our parameter settings We used the default parameter settings.

Software availability `http://raptorx.uchicago.edu/`

2.21 Mutagenetic: mutagenetic tree mixture model [21] (2005)

Introduction Mutagenetic tree models have been designed to investigate the accumulation of drug resistance-associated mutations in HIV-1 proteins [21]. It was shown that mutagenetic tree mixture models could identify many parallel or confluent mutation pathways using sequence datasets of HIV-1 protease [21]. This method has also been applied to the field of tumor development [47, 48].

Methodology The mutagenetic tree model is built using a set of directed weighted trees. This model can approximate the joint probability distribution consisting of a set of mutational events. Based on a similar methodology proposed in the mixed tree probabilistic models [49], EM algorithm can be used to maximize the log-likelihood function. Suppose a mutagenetic tree is denoted as $T = (V, E)$ with the set of vertices V and the set of edges E, $x = \{x_i, ..., x_N\}$ represents the mutation pattern. Given one mutagenetic tree model T, the likelihood of a pattern x in T can be modeled as:

$$L(x|T) = \prod_{e \in E(V(x))} P(e) \times \prod_{e \in E(V - V(x))} (1 - P(e)) \tag{35}$$

The mixed tree model is defined by $M = \sum_{k=1}^{K} \alpha_k T_k$, where $\alpha_k \in [0, 1]$ is the weight of the k^{th} tree T_k and $\sum_{k=1}^{K} \alpha_k = 1$. The likelihood of the trained mutagenetic tree M is defined as:

$$f(x_1, ..., x_N|M) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} \alpha_k L(x_i|T_k) \tag{36}$$

The parameters α_k and T_k are thereafter optimized by the EM algorithm [49].

Our parameter settings For each input sequence dataset, one mutagenetic tree was created using default settings.

Software availability `http://mtreemix.bioinf.mpi-inf.mpg.de/`.

2.22 BN: Bayesian network [22] (2007)

Introduction Bayesian networks (BNs) have been used to model mutational pathways in HIV-1 proteins [22, 50, 51]. It was shown that Bayesian networks could be useful for improving drug resistance predictions [52].

Methodology Bayesian networks are trained based on methods adapted from the Bright software [53]. Given a MSA input, this method maximizes the posterior probability of Bayesian networks whose variables are residues or therapies. The robustness of Bayesian networks is examined by a non-parametric bootstrap resampling using 100 replicates. In a consensus Bayesian network, edges between variable nodes are considered as robust if their bootstrap supports are above 65%. Different amino acids at the same position may cluster together due to the presence of strong antagonistic effects.

Our parameter settings We used the default parameters as indicated in [22] (bootstrap resampling: 100 replicates, bootstrap support: 65%). In order to compare results of Bayesian networks with other coevolution methods, we extracted position pairs of adjacent variables in the trained Bayesian networks.

Software availability `https://code.google.com/p/bright/`.

2.23 Spidermonkey [23] (2008)

Introduction In the software platform Spidermonkey, coevolving positions are modeled using Bayesian networks trained using reconstructed ancestral sequences from a phylogenetic tree [23]. More specifically, Spidermonkey uses MCMC-based algorithms to model conditional dependencies between the non-synonymous positions in Bayesian networks. The advantage of this method relies on the fact that Bayesian networks can model high order interactions using ancestral sequences, for coevolving positions may have high-order interactions in protein families [23]. However, the convergence of MCMC in the process of Bayesian network training requires large sequence datasets and a heavy demand of compu-

tation power. As a compromise between computing power and prediction accuracy, Spidermonkey allows for at most two parent nodes of variables in Bayesian networks [23].

Methodology Given a sequence dataset, Spidermonkey estimates a substitution model and reconstructs ancestral sequences in the phylogenetic tree using a strategy of maximum-likelihood optimization. The statistical dependencies between protein positions in Bayesian networks are modeled using a MCMC-based algorithm [23].

Our parameter settings In our analysis, we used parameter settings as follows: (1) nucleotide model 012345, MG94x; (2) treatment of ambiguities: averaged; (3) the number of positions with substitution values: default, (5) maximum parents: 2; (6) the number of MCMC chains: default (100000); (7) the number of burn-in steps before the main chain: default (10000); (8) the number of ancestral samples: default (100).

Software availability `http://www.hyphy.org/w/index.php/Main_Page`.

2.24 CTMP: continuous time Markov process [24] (2007)

Introduction This study proposed a continuous-time Markov process model augmented with the phylogenetic information [24]. To identify sequence coevolution in different protein families, the model was applied to screen all position pairs of inter- and intra-domains in the protein families [24]. The majority of coevolving protein domains was found near functionally important positions, providing an interesting information of protein structural and functional constraints in the sequence coevolution [24].

Methodology Four steps are performed to measure the residue coevolution. Firstly, the matrix of coevolutionary rates corresponding to amino acid changes is obtained by reweighting the independent coevolutionary rates. Secondly, sequences from different protein domains are mapped to the leaves of phylogenetic trees shared with the same topology. Thirdly, the log-likelihood ratio is measured by the likelihood of observed sequences in the coevolution model compared to the null model. By doing so, the joint probability of residue positions is approximated by the singlet and pairwise terms of aligned positions among all states of

internal nodes in the phylogenetic tree. The CTMP model is simplified as:

$$P(x_1(t),...,x_n(t)|x_1(0),...,x_n(0)) = \frac{\prod\limits_{x_i - x_j \in \pi} P(x_i(t), x_j(t)|x_i(0), x_j(0))}{\prod\limits_{i=1}^{n} P(x_i(t)|x_i(0))^{d_i - 1}} \quad (37)$$

Where $x_i(t)$ is the sequence composition at the i^{th} position with the sampling time t given the phylogenetic tree π and the MSA input. The position pair $x_i(t), x_j(t)$ is observed on the leaves of the phylogenetic tree π.

Lastly, the false positive rates of coevolving position pairs are evaluated by multiple hypothesis tests using the simulated datasets.

Our parameter settings We used maximum 500 sequences for our CTMP analysis due to limited computation power. The maximum-likelihood phylogenetic trees were trained using RAxML V7.0.4. The threshold on the fraction of sequences coevolving with non-overlapping states was set to 1, as well as the threshold on the fraction of conserved sequences. The penalty parameter (ϵ) of the CTMP model was set to 0.25 according to the manual.

Software availability `http://www.stat.sinica.edu.tw/chyeang/`.

2.25 CoMap [25] (2011)

Introduction CoMap V1.4.1 uses Markov models to identify residue coevolution based on the phylogenetic tree [25]. This model takes into account the uncertainty of ancestral states and among-site rate variations given the phylogenetic tree inputs [25]. As an advantage, CoMap can work on both nucleotide and amino acid sequence datasets. Using a ribosomal RNA dataset including 79 bacteria species, this method identified more than 95% intra-protein predictions based on protein contact maps [25].

Methodology Firstly, for each residue position, CoMap creates a substitution vector which contains posterior estimates of the substitutions at each branch given a phylogenetic tree. By accounting for position variations, this substitution vector is defined as:

$$v_{i,b} = \sum_{c} \sum_{x_p} \sum_{x_q} n_{x_p,x_q}(t \cdot r_c) \times P(x_p, x_q, r_c|D_i, \Theta) \quad (38)$$

Where D_i is the i^{th} position of the MSA input D, a is the number of internal nodes in the phylogenetic tree and b is a branch in the phylogenetic tree, r_c is the rate of class c, Θ is the set of parameters including branch lengths, substitution matrices and rate distribution parameters. $n_{x_p,x_q}(t \cdot r_c)$ is the conditional observation of substitutions expected on the branch with its branch length t and the states x_p, x_q.

Secondly, the Pearson's correlation coefficient between two substitution vectors is:

$$\rho_{i,j} = \frac{Cov(V_i, V_j)}{\sigma(V_i)\sigma(V_j)} \tag{39}$$

Where $V_i = (v_{i,1}, ..., v_{i,b}, ..., v_{i,m})$ is obtained in the first step and $\sigma(V_i)$ is the standard deviation of V_i.

Thirdly, to show the statistical significance, p-values are measured by comparing the conditional observations of substitutions with the expectation of the null hyphothesis of independence. The null distribution is estimated by simulating 100000 independent pairs [25].

Our parameter settings We used pairwise analysis to calculate the compensation coefficient for each position pair in CoMap V1.4.1 [54]. Due to the limitation of our computation power, the number of sequences in the input datasets was restrained to be less than 500 sequences. In the first round, CoMap reported varied positions in the maximum likelihood phylogenetic tree. We thereafter removed positions reported with infinite maximum likelihood according to the software manual. In the second round, the maximum-likelihood phylogenetic trees were prepared using RAxML V7.0.4. As suggested in the software manual, we used the following parameters: (1) nijt_aadist.sym=no, (2) aadist.type=grantham, (3) statistic=Compenstation, (4) model=LG08, (5) statistic.null=yes, (6) statistic.null.compute_pvalue=yes, and (7) statistic.null.nb_rep_CPU=8. The coevolving residues were collected if their p-values were less than 0.05.

Software availability `http://gna.org/projects/comap`.

2.26 GREMLIN: generative regularized models of proteins [26] (2013)

Introduction GREMLIN was originally proposed to learn an undirected probabilistic graphical model of the amino acid compositions given the inputs of MSA [55]. GREMLIN outperformed hidden Markov models using the datasets of 71 protein families extracted from the PFAM database [55]. As sequence-based methods usually require a large amount of sequences to achieve robust predictions, this paper contributes to incorporate prior information

on residue pairs so that fewer sequences are needed for robust coevolution predictions [26]. Performance of GREMLIN was compared to MI, PSICOV, DCA, plmDCA and MIc using a large sequence dataset with 329 protein families.

Methodology Markov random field is used to model the probability distribution given a set of independent sequences $X = \{X^1, X^2, ..., X^N\}$ in the MSA input. Due to the intractable computation of the global maximum likelihood function, the pseudo-likelihood function is proposed for the efficient approximation to make the problem solvable. Specifically, the pseudo-likelihood is defined as:

$$pll(\Phi) = \frac{1}{N} \sum_{X^i \in X} \sum_{j=1}^{N} \left[\log \phi_j(X_j^i) + \sum_{k \in ne(V_j)} \log \phi_{jk}(X_j^i, X_k^i) - \log Z_j \right] \tag{40}$$

Where X_j^i is the residue at the j^{th} position of the i^{th} sequence given the MSA input, ϕ_j is the potential function of the position j, ϕ_{jk} is the potential function for the edge $j-k$, Z_j is a local normalization constant, $ne(V_j)$ is the set of vertices connected with the node j in the undirected graphical model.

For both structure learning and parameter estimation, the L1 regularization is used for the optimization based on the projected gradients.

$$max_{\Phi,\alpha} \quad pll(\Phi) - \lambda_{node} \sum_{s=1}^{N} ||V^s||_2^2 - \lambda_{edge} \sum_{s=1}^{N} \sum_{t=s+1}^{N} \alpha_{st} \tag{41}$$

$$subject \quad to: \quad \forall (1 \leq s < t \leq N) : \alpha_{st} \geq ||W^{st}||_2 \tag{42}$$

Where λ_{node} and λ_{edge} are regularization parameters that determine the weights of the penalty level. α_{st} is the differentiable proxy of $||W^{st}||_2$, which solves the calculation using a smooth convex optimization.

Our parameter settings We used the default parameter settings.

Software availability `http://openseq.org/`.

3 Implementation and software manual

Several sequence-based methods (e.g. SCA, LogR, DCA, GREMLIN, ZERS, NBZPX2) were designed using Matlab, we thus decided to build up the ensemble coevolution system

in Matlab by integrating the available methods and implementing 7 methods without public sources. We newly implemented 7 methods in our toolbox including ASC, APC, Complementary, NCPS, PCC, PhysicoMI and RCW. The other methods designed in non-Matlab platform were requested from the authors or downloaded from public sources (see Table 1).

Due to the copy right issue, users who wish to test methods implemented in other publications need to install software independently. Nevertheless, our toolbox offers users the platform to prepare the input files, the command lines (if any), the extraction of output results and the assembly of all 27 methods integrated in our system. Currently, the toolbox V0.1 has only been tested in Linux (Ubuntun 12.04, 64-bit). If system comparability or other issues were encountered, please write an email to liguangdi.research@gmail.com to obtain the latest version.

References

1. Brandman R, Brandman Y, Pande VS (2012) Sequence coevolution between rna and protein characterized by mutual information between residue triplets. PloS one 7: e30022.
2. Dunn SD, Wahl LM, Gloor GB (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24: 333–340.
3. Gouveia-Oliveira R, Pedersen AG, et al. (2007) Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. Algorithms Mol Biol 2: 12.
4. Little DY, Chen L (2009) Identification of coevolving residues and coevolution potentials emphasizing structure, bond formation and catalytic coordination in protein evolution. PloS one 4: e4762.
5. Tillier ER, Lui TW (2003) Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. Bioinformatics 19: 750–755.
6. Ackerman SH, Tillier ER, Gatti DL (2012) Accurate simulation and detection of coevolution signals in multiple sequence alignments. PloS one 7: e47108.
7. Gao H, Dou Y, Yang J, Wang J (2011) New methods to measure residues coevolution in proteins. BMC bioinformatics 12: 206.
8. Kalinina OV, Oberwinkler H, Glass B, Kräusslich HG, Russell RB, et al. (2012) Computational identification of novel amino-acid interactions in hiv gag via correlated evolution. PloS one 7: e42468.
9. Lee BC, Kim D (2009) A new method for revealing correlated mutations under the structural and functional constraints in proteins. Bioinformatics 25: 2506–2513.
10. Reynolds KA, McLaughlin RN, Ranganathan R (2011) Hot spots for allosteric regulation on protein surfaces. Cell 147: 1564–1575.

11. Halperin I, Wolfson H, Nussinov R (2006) Correlated mutations: Advances and limitations. a study on fusion proteins and on the cohesin-dockerin families. Proteins: Structure, Function, and Bioinformatics 63: 832–845.

12. Ashkenazy H, Kliger Y (2010) Reducing phylogenetic bias in correlated mutation analysis. Protein Engineering Design and Selection 23: 321–326.

13. Burger L, van Nimwegen E (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS computational biology 6: e1000633.

14. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences 108: E1293–E1301.

15. Jones DT, Buchan DW, Cozzetto D, Pontil M (2012) Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28: 184–190.

16. Cheng J, Baldi P (2007) Improved residue contact prediction using support vector machines and a large feature set. BMC bioinformatics 8: 113.

17. Tegge AN, Wang Z, Eickholt J, Cheng J (2009) Nncon: improved protein contact map prediction using 2d-recursive neural networks. Nucleic Acids Research 37: W515–W518.

18. Eickholt J, Cheng J (2012) Predicting protein residue–residue contacts using deep networks and boosting. Bioinformatics 28: 3066–3072.

19. Di Lena P, Nagata K, Baldi P (2012) Deep architectures for protein contact map prediction. Bioinformatics 28: 2449–2457.

20. Wang Z, Xu J (2013) Predicting protein contact map using evolutionary and physical constraints by integer programming. Bioinformatics 29: i266–i273.

21. Beerenwinkel N, Rahnenführer J, Däumer M, Hoffmann D, Kaiser R, et al. (2005) Learning multiple evolutionary pathways from cross-sectional data. Journal of Computational Biology 12: 584–598.

22. Deforche K, Silander T, Camacho R, Grossman Z, Soares M, et al. (2006) Analysis of hiv-1 pol sequences using bayesian networks: implications for drug resistance. Bioinformatics 22: 2975–2979.

23. Poon AF, Lewis FI, Frost SD, Pond SLK (2008) Spidermonkey: rapid detection of co-evolving sites using bayesian graphical models. Bioinformatics 24: 1949–1950.

24. Yeang CH, Haussler D (2007) Detecting coevolution in and among protein domains. PLoS computational biology 3: e211.

25. Dutheil J, Pupko T, Jean-Marie A, Galtier N (2005) A model-based approach for detecting coevolving positions in a molecule. Molecular biology and evolution 22: 1919–1928.

26. Kamisetty H, Ovchinnikov S, Baker D (2013) Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. Proceedings of the National Academy of Sciences 110: 15674–15679.

27. Polikar R (2006) Ensemble based systems in decision making. Circuits and Systems Magazine, IEEE 6: 21–45.

28. Polikar R (2012) Ensemble learning. In: Ensemble Machine Learning, Springer. pp. 1–34.

29. Troć M, Unold O (2010) Self-adaptation of parameters in a learning classifier system ensemble machine. International Journal of Applied Mathematics and Computer Science 20: 157–174.

30. Gao Y, Huang JZ, Wu L (2007) Learning classifier system ensemble and compact rule set. Connection Science 19: 321–337.

31. Bacardit J, Krasnogor N (2008) Empirical evaluation of ensemble techniques for a pittsburgh learning classifier system. In: Learning Classifier Systems, Springer. pp. 255–268.

32. Gama J, Brazdil P (2000) Cascade generalization. Machine Learning 41: 315–343.

33. Saha I, Zubek J, Klingstrom T, Forsberg S, Wikander J, et al. (2014) Ensemble learning prediction of protein-protein interactions using proteins functional annotations. Molecular BioSystems .

34. Yang J, Jang R, Zhang Y, Shen HB (2013) High-accuracy prediction of transmembrane inter-helix contacts and application to gpcr 3d structure modeling. Bioinformatics 29: 2579–2587.

35. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: Using pseudolikelihoods to infer potts models. Physical Review E 87: 012707.

36. Cover TM, Thomas JA (2012) Elements of information theory. Wiley-interscience.

37. Bremm S, Schreck T, Boba P, Held S, Hamacher K (2010) Computing and visually analyzing mutual information in molecular co-evolution. BMC bioinformatics 11: 330.

38. Bielza C, Li G, Larranaga P (2011) Multi-dimensional classification with bayesian networks. International Journal of Approximate Reasoning 52: 705–727.

39. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286: 295–299.

40. Süel GM, Lockless SW, Wall MA, Ranganathan R (2002) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nature Structural & Molecular Biology 10: 59–69.

41. Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R (2005) Natural-like function in artificial ww domains. Nature 437: 579–583.

42. Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein sectors: evolutionary units of three-dimensional structure. Cell 138: 774–786.

43. Bartlett GJ, Taylor WR (2008) Using scores derived from statistical coupling analysis to distinguish correct and incorrect folds in de-novo protein structure prediction. Proteins: Structure, Function, and Bioinformatics 71: 950–959.

44. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N (2001) Residue frequencies and pairing preferences at protein–protein interfaces. Proteins: Structure, Function, and Bioinformatics 43: 89–102.

45. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein–protein interaction by message passing. Proceedings of the National Academy of Sciences 106: 67–72.

46. Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9: 432–441.

47. Rahnenführer J, Beerenwinkel N, Schulz WA, Hartmann C, Von Deimling A, et al. (2005) Estimating cancer survival and clinical outcome based on genetic tumor progression scores. Bioinformatics 21: 2438–2446.

48. Desper R, Jiang F, Kallioniemi OP, Moch H, Papadimitriou CH, et al. (1999) Inferring tree models for oncogenesis from comparative genome hybridization data. Journal of Computational Biology 6: 37–51.

49. Meila M, Jordan MI (2001) Learning with mixtures of trees. The Journal of Machine Learning Research 1: 1–48.

50. Deforche K, Camacho R, Grossman Z, Silander T, Soares M, et al. (2007) Bayesian network analysis of resistance pathways against hiv-1 protease inhibitors. Infection, Genetics and Evolution 7: 382–390.

51. Deforche K, Camacho RJ, Grossman Z, Soares MA, Van Laethem K, et al. (2008) Bayesian network analyses of resistance pathways against efavirenz and nevirapine. Aids 22: 2107–2115.

52. Deforche K, Camacho R, Van Laethem K, Lemey P, Rambaut A, et al. (2008) Estimation of an in vivo fitness landscape experienced by hiv-1 under drug selective pressure useful for prediction of drug resistance evolution during treatment. Bioinformatics 24: 34–41.

53. Myllymäki P, Silander T, Tirri H, Uronen P (2002) B-course: A web-based tool for bayesian and causal data analysis. International Journal on Artificial Intelligence Tools 11: 369–387.

54. Dutheil J, Galtier N (2007) Detecting groups of coevolving positions in a molecule: a clustering approach. BMC evolutionary biology 7: 242.

55. Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ (2011) Learning generative models for protein fold families. Proteins: Structure, Function, and Bioinformatics 79: 1061–1078.

Chapter 6

HIV-1 Gag-protease coevolution network

"If you can believe it, the mind can achieve it."

— Ronnie Lott

This chapter is adapted from my article:

Guangdi Li, Jens Verheyen, Kristof Theys, Andrea-Clemencia Pineda-Peña, Ricardo Khouri, Kristel Van Laethem, Jan Ramon and Anne-Mieke Vandamme. HIV-1 Gag-protease coevolution networks.

I proposed the idea, performed the sequence analysis and drafted the manuscript. The improvement of the paper was supported with substantial advices and corrections from my coauthors. I sincerely thank Fossie Ferreira and Supinya Piampongsant for technical assistance and valuable contributions to the analysis.

6.1 Summary

HIV-1 infected patients who failed protease inhibitor (PI) treatment can harbor viral strains with Gag mutations. Yet, the coevolution between HIV-1 protease and its substrate Gag under PI selective pressure is still not fully understood. This study investigates the coevolution network between HIV-1 subtype B Gag and protease under PI selective pressure, and to evaluate its impact on current Gag inhibitors and HIV-human protein interactions. We sequenced the Gag and protease regions from 531 patients in our Leuven cohort, and extracted 514 genomic and 3171 Gag-protease sequences from public resources. Using an ensemble coevolution method, 31 Gag-protease coevolving pairs of amino acid positions were predicted, 13 of them were evaluated as true positives by *in vitro* and *in vivo* data. All predicted coevolving pairs harbored Gag-protease mutations whose prevalence significantly differed between the PI-resistant and the PI-susceptible sequence populations (p-value<0.01). HIV-1 Gag-protease coevolution networks consisted of 29 Gag positions associated with 30 protease positions. Among these 29 PI-associated Gag positions, 13 positions also interacted with human proteins (e.g. APOBEC3G, Cyclophilin A); while only 6 positions interacted with 8% of 50 Gag experimental inhibitors. Moreover, 62.1% of 29 PI-associated Gag positions and 80% of 10 Gag mutations identified in our Leuven cohort were located within either Gag cleavage sites or Gag C terminus (positions: 362-500). Genome-wide sequence analysis independently showed that the amino acid diversity of drug-targeted regions and Gag cleavage sites and C terminus was significantly associated with drug selective pressure (p-value<0.01). HIV-1 can escape PI selective pressure by selecting protease drug resistance and Gag mutations mainly located at Gag cleavage sites and C terminus. HIV-1 Gag-protease coevolution takes place in the presence of human proteins, but it is unlikely to affect Gag experimental inhibitors.

6.2 Introduction

HIV is a retrovirus that produces up to 10 billion viral particles within a single patient per day. Immature HIV virions released from human cells are transformed into mature virions after the protease-mediated proteolytic cleavage of Gag and Gag-Pol polyproteins. This process leads to the morphogenesis of structural proteins (matrix,

capsid, nucleocapsid (NC), p6), enzymatic proteins (protease, reverse transcriptase, integrase) and spacer peptides (p1, p2) [1] (**Figure S 6.1**). HIV protease recognizes Gag and Gag-Pol cleavage sites through conserved structural conformations [2]. Proteolytic processing proceeds in a relatively strict order at a rate that varies between cleavage sites, and is susceptible to the surrounding contexts [3]. Since this proteolytic mechanism plays a major role during the viral maturation, protease represents an attractive anti-HIV drug target [4]. Many FDA-approved protease inhibitors (PIs) have been designed to competitively bind the protease catalytic pocket [5]. However, PI drug resistance can be caused by the evolutionary flexibility of HIV-1 protease and protease substrates [6].

Position-specific coevolution is the reciprocal evolutionary process of amino acids, often observed within interacting protein positions [7, 8]. While position-specific coevolution on the conserved surface of transient protein-protein interactions is relatively slow [9, 10], HIV protease and its substrate Gag may coevolve much faster due to the high mutation rate via the error-prone reverse transcription [11]. Since PIs target protease positions that interact with protease substrates, amino acid substitutions in protease and Gag selected for in response to these drugs may cause conformational changes that affect binding affinity and impair viral replication [6]. Residues surrounding Gag cleavage sites may compensate for these conformational changes to maintain viral productivity [12].

Protease structures have been crystallized with peptide analogs of protease substrates (approximately 10 amino acids, **Figure S 6.1**). Weak bindings [13], low interaction energies [14] and reduced van der Waals contacts [15] can affect the interactions between drug-resistant protease mutants and natural substrate peptides. NC-p1 peptide substrates in the presence of Gag mutation A431V bind to protease V82A mutant more tightly than wild-type protease, because of increased hydrogen bonds and van der Waals contacts [16]. Nevertheless, a macromolecular Gag-protease complex has not been crystallized owing to the flexible nature of Gag inter-domain links [17]. Taken together, current structural evidence has shown that some cleavage site mutations (CSMs) in Gag can compensate for the reduced Gag-protease interactions [8]. However, lack of structural information on the macromolecular Gag-protease

complex imposes a challenge to explore the coevolution between the full-length Gag and protease.

Previous clinical studies have analyzed different patient cohorts (e.g. AREVIR [18], NARVAL [19], RESINA [20]) to identify HIV-1 Gag mutations (e.g. A431V, I437V, L449V) emerging during PI-based regimen, mostly in the presence of protease mutations (e.g. M46I, L76V, V82A). In the absence of protease mutations, Gag mutations may cause partial PI resistance, but these occurrences are low [21, 22]. As HIV-1 Gag mutations are associated with PI drug resistance, the coevolution between the full-length Gag and protease warrants a comprehensive investigation.

This study investigates the coevolution between the full-length Gag and protease of HIV-1 subtype B under PI selective pressure. HIV-1 Gag-protease coevolution networks were constructed using sequence-based methods that detect coevolving pairs of amino acid positions in the full-length Gag and protease. To evaluate our networks, we reviewed *in vitro* and *in vivo* studies to collect a comprehensive list of documented Gag-protease mutation patterns associated with PI susceptibility. We then presented longitudinal data of our Leuven patient cohort to show Gag substitutions emerging during PI treatment. We also explored the potential impact of Gag-protease coevolution on Gag inhibitors and HIV-human protein interactions.

6.3 Materials and Methods

Dataset of PI-susceptible and PI-resistant Gag-protease sequences

We retrieved 11812 nucleotide sequences of HIV-1 subtype B Gag and protease proteins from the Los Alamos HIV database (parameters: HXB2 nucleotide region: 790-2550, minimum nucleotide length: 300, one sequence per patient). We thereafter selected 9320 sequences containing the full-length protease and the partial or full-length Gag. Sequences were then aligned against the HXB2 reference and manually curated using Seaview V4.3 [23]. Sequence subtypes were assessed using the subtyping tools of Rega V3.0 [24] and COMET V1.0 (http://comet.retrovirology.lu/). To improve sequence quality, we excluded sequences with ≥99% similarity, hypermutations, stop codons or discordant subtype classifications [25]. This procedure resulted in the Gag-protease sequence dataset containing 3171 HIV-1 subtype B Gag and protease sequences.

For these 3171 sequences, patient treatment information was collected from the original publications. Approximately 90% of Gag-protease sequences were sampled from treatment-naïve patients, while detailed treatment information of PI-treated patients (e.g. drug combination, therapy duration, treatment outcome) was largely lacking. For this reason, we used the drug resistance interpretation algorithms, HIVdb V6.0 [26] and Rega V9.1 [27], to assess PI susceptibility levels and to deduce assumed PI exposure. We defined two subsets of Gag-protease sequences based on the predicted PI activity. The Gag-protease PI-susceptible dataset contained 1820 Gag and protease sequences, which were sampled from PI-naïve patients and were interpreted as susceptible to all PIs by both interpretation algorithms. The Gag-protease PI-resistant dataset contained 434 Gag and protease sequences, which were predicted to be (partially or fully) resistant against at least one common PI by both algorithms. For our Gag-protease coevolution analysis, we further selected full-length Gag and protease sequences, leading to 759 PI-susceptible and 168 PI-resistant sequences in the so-called "full-gagpro-susceptible" and "full-gagpro-resistant" datasets, respectively. Merging these two latter datasets led to the "full-gagpro" sequence dataset. Given the full-gagpro-susceptible sequence dataset, intra-subtype sequence diversity was calculated at each amino acid position of Gag and protease using our method described previously [25].

Amino acid diversity in the HIV-1 B full-length genome

We retrieved 672 full-length subtype B genomic sequences from the Los Alamos HIV database (one genome per patient). Following similar procedures described above (sequence alignment, sequence quality control, drug susceptibility tests), we obtained a subtype B genomic dataset including 94 drug-resistant and 420 drug-susceptible genomic sequences, respectively. Drug resistance interpretation algorithms HIVdb V6.0 [26] and Rega V9.1 [27] were used to perform drug susceptibility tests on three drug classes: PI, RTI (reverse transcriptase inhibitor) and INI (integrase inhibitor). Next, we concatenated amino acid sequences of 15 HIV-1 proteins in the full-length genome. Using bootstrap resampling with 1000 replicates, pairwise amino acid diversity at each position was calculated using drug-susceptible and drug-resistant genomic sequence datasets [25]. Mann–Whitney U test was then performed to compare the distributions of amino acid diversity. A significant difference was detected if a p-value was lower than 0.05.

Sequence and treatment information of Leuven longitudinal dataset

We collected 637 Gag and protease sequences from 531 subtype B infected patients, attending the University Hospital of Leuven. Sequences were obtained during clinical follow-up between 1996 and 2013. Our sequencing protocol and quality control procedures were described previously [25, 28, 29]. We extracted patient information (history of antiretroviral treatment, sampling time, viral load) from our Leuven database [30]. Of these 637 Gag-protease sequences, all contained the protease region, 22 had the full-length Gag and the remaining 615 sequences contained the Gag C terminus. To identify amino acid substitutions that emerged during PI-based regimen, we prepared the longitudinal dataset of 44 patients who had taken at least one PI for more than four weeks, and had more than one Gag-protease sequences sampled during PI treatment.

Documented Gag-protease mutation patterns in literature

We reviewed journal articles to collect the *in vitro* and *in vivo* data of Gag-protease mutation patterns documented in HIV-1 subtype B. We searched for English articles in PubMed published between January 1983 and September 2013 using the keywords ("HIV Gag mutation", "HIV Gag protease", "HIV protease mutations Gag", "HIV Gag evolution", "HIV protease cleavage"). English articles were also reviewed if they were referenced in literature or published full-length Gag sequences. **Table S 6.1** summarizes the documented Gag-protease mutation patterns, *in vivo* data (patient treatments, sampling size, study cohort) and *in vitro* data (viral replication capacity, drug susceptibility to the following 8 PIs: TPV, SQV, APV, DRV, NFV, IDV, RTV, ATV).

HIV-1 PDB and protein secondary structures

From the RCSB Protein Data Bank (www.pdb.org), we collected the PDB data of HIV-1 matrix, capsid, p2, nucleocapsid, p6, protease and PI-protease complexes. The quality of PDB data was assessed using PDBREPORT [31] (default parameters). We used PSIPRED [32] and 2Struc [33] to determine protein secondary structures (e.g. alpha-helix, beta-strand, random-coil); consensus results were used for our analysis.

Gag experimental inhibitors, HIV-human protein interactions and human CD4/CD8/antibody epitope positions

Three datasets were prepared as follows. (A) We retrieved 137 drug binding positions of 50 Gag experimental inhibitors from our recent study [25]. (B) From the NCBI HIV human protein interaction database [34], we collected information on direct interactions between 46 human proteins and 4 HIV-1 Gag proteins (matrix, capsid, nucleocapsid, p6). **Table S 6.2** summarizes the information of HIV-human protein interaction. (C) We collected human CD4 T cell and antibody epitopes from the HIV Los Alamos database [35]. For human CTL/CD8 T cell epitopes, we included the A-list which comprised the best-defined CTL/CD8 epitopes described by Llano et al [35] (**Table S 6.3**).

Ratio of synonymous and non-synonymous substitution rate (dN/dS)

Given the full-gagpro-resistant sequence dataset, we constructed a maximum-likelihood phylogenetic tree using FastTree V2.1 [36] (parameters: continuous gamma distribution, fully optimized generalized time-reversible (GTR) model). Provided with the constructed phylogenetic tree and the full-gagpro-resistant sequence dataset, we then applied HyPhy V2.1.0 [37] to estimate the ratio of non-synonymous and synonymous rates (dN/dS). We employed the single likelihood ancestor counting (SLAC) model with the optimized GTR [37]. Ambiguous nucleotides were resolved by averaging over all possible states for the ancestral sequence reconstruction [37]. Statistical significance of dN/dS was measured by the continuous extension of binomial distributions (significance level: 0.01) [37].

Prevalence of HIV-1 subtype B Gag-protease mutations

We calculated the prevalence of Gag-protease mutations in the PI-susceptible (n=1820) and PI-resistant (n=434) sequence datasets described above. Two-tailed Fisher's exact tests were performed to examine the statistical significance of mutation prevalence differed between these two sequence datasets. The adjusted p-values were obtained using multiple testing correction via the false discovery rate approach [38]. Odds-ratio (OR) test was used to determine the odds of a Gag-protease mutation being in the PI-resistant sequence dataset relative to the odds of the mutation being in the PI-susceptible sequence dataset. Specifically, for a Gag mutation X and a protease mutation Y, OR is defined as: $OR = [P_{\mathrm{Res}} / (1 - P_{\mathrm{Res}})] / [P_{Sus} / (1 - P_{Sus})]$, where $P_{\mathrm{Res}}(X,Y)$ and $P_{\mathrm{Sus}}(X,Y)$ represent the probability of Gag-protease mutation (X,Y) in the full-gagpro-resistant and the full-gagpro-susceptible datasets, respectively. Zero values in the OR

contingency tables were replaced with 0.5 to avoid an impossible infinity [39]. The 95% confidence interval of OR was determined using standard procedures [40].

Construction of Gag-protease coevolution networks

Provided with the full-gagpro-resistant and full-gagpro-susceptible sequence datasets, we used our ensemble coevolution system (ECS) to predict coevolving pairs of amino acid positions, so-called coevolving pairs. Given an input of multiple sequence alignments (MSAs), ECS detects coevolving pairs in a three-step process (**Figure 6.1**A): (1) Sequence datasets are prepared by bootstrap resampling with 100 replicates. (2) Resampled sequence datasets are imported into four sequence-based methods (CMPro [41], NCPS [42], PhyCMAP [43], RCW [44]), each of which estimates a statistical score for a coevolving pair. (3) The combiner assembles predictions of coevolving pairs from the four sequence-based methods, treated with equal weights. Thereafter, coevolving pairs with their statistical scores are ranked and exported as outputs.

A four-step process constructed HIV-1 Gag-protease coevolution networks (**Figure 6.1**B). (1) The full-gagpro-resistant (so-called D1) and the full-gagpro-susceptible (so-called D2) sequence datasets were imported into ECS for the prediction of coevolving pairs. The outputs of top-ranked coevolving pairs predicted from D1 and D2 were collected into the output datasets O1 and O2, respectively. Predicted coevolving pairs between a Gag position and a protease position, termed Gag-protease coevolving pairs, were separated from those coevolving pairs of positions within individual Gag and protease proteins.

(2) Differential predictions between O1 and O2, denoted as O1-O2, were distinguished to identify top-ranked Gag-protease coevolving pairs that were associated with genotypic PI resistance. The threshold of top-ranked predictions was set to $3 \times L$, where L was the number of amino acid positions in the MSA (L=599 in our datasets).

(3) Top-ranked Gag-protease coevolving pairs were considered as significant if they contained any Gag-protease mutation that met two conditions: (a) Prevalence of the Gag-protease mutation was higher than 1% in the full-gagpro-resistant dataset. (b) Prevalence of the Gag-protease mutation differed significantly between the full-gagpro-resistant and the full-gagpro-susceptible datasets (p-value < 0.01).

(4) HIV-1 Gag-protease coevolution networks were constructed. A top-ranked Gag-protease coevolving pair was considered as a true positive if it contained one of the Gag-protease mutation patterns documented in the *in vitro* and *in vivo* datasets (**Table S 6.1**). Apart from the true positives, the unconfirmed predictions and the unpredicted positives (documented Gag-protease mutation patterns but not predicted) were also mapped.

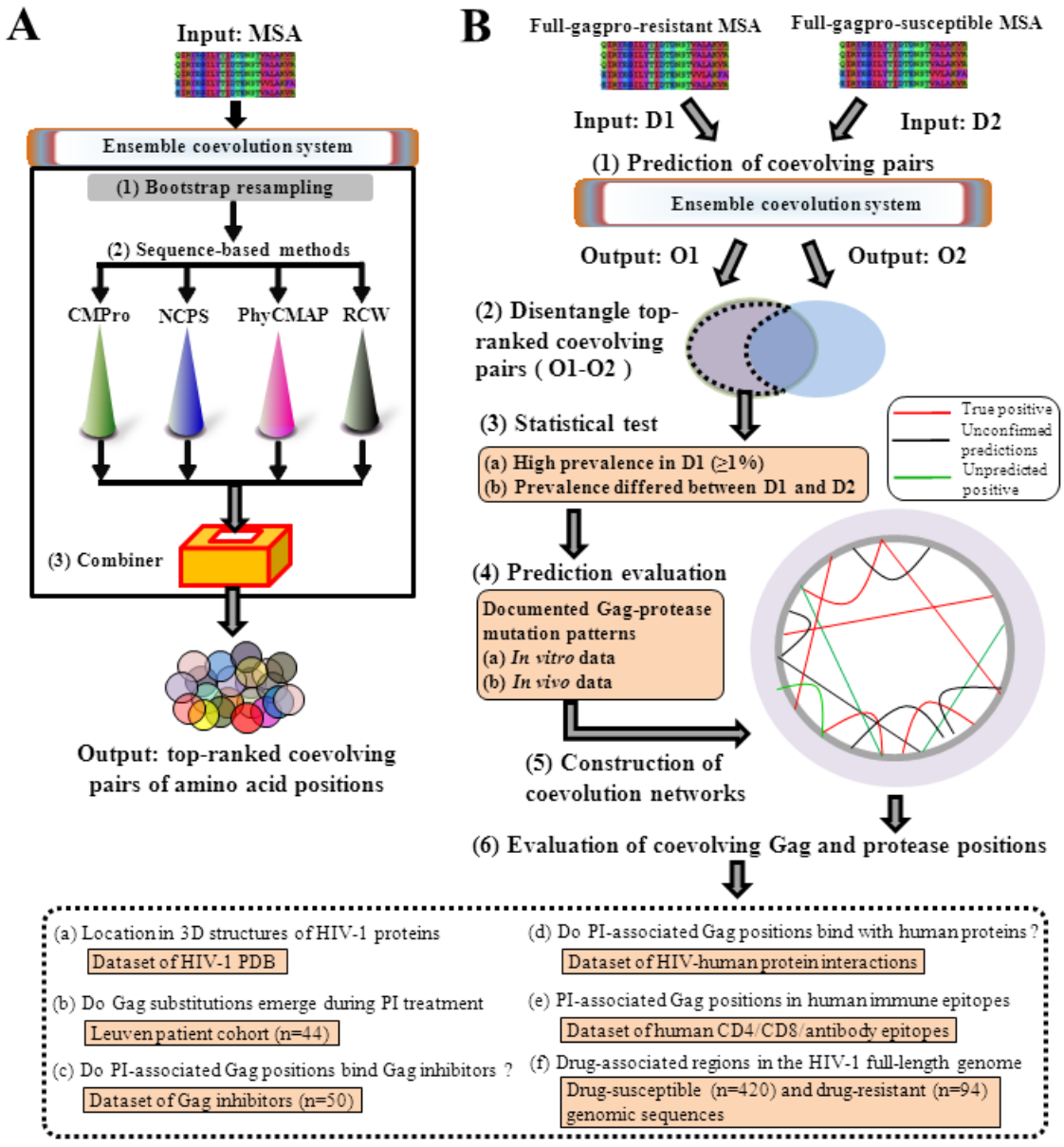


Figure 6.1: (A) Schematic view of ensemble coevolution system (ECS). Three major components are illustrated: (1) Input data, one MSA. (2) Bootstrap resampling, 100 sequence datasets are generated by the bootstrap resampling with 100 replicates. (3) Sequence-based methods, resampling sequence datasets are imported into four sequence-based methods (CMPro [41], NCPS [42], PhyCMAP [43], RCW [44]); each predicts a coevolution score for a coevolving pair of amino acid positions. (4)

Combiner, for each coevolving pair predicted by individual sequence-based methods, coevolution scores are averaged over 100 resampling sequence datasets and normalized between 0 and 1. Given the equal weights of individual sequence-based methods, coevolving pairs are ranked based on normalized coevolution scores.
(B) Workflow of coevolution network construction. Five steps are processed. (1) Predictions of coevolving pairs, the PI-resistant and PI-susceptible datasets are imported into ECS (see (A)) for the detection of coevolving pairs of amino acid positions. The prediction outputs from PI-resistant and PI-susceptible datasets are denoted as O1 and O2, respectively. (2) Disentangle top-ranked coevolving pairs, top-ranked predictions from the output O1 are disentangled from the output O2. (3) Statistical test, top-ranked predictions in the disentangled dataset O1-O2 are further selected if they have a high prevalence in D1 (≥1%) and a prevalence differed between D1 and D2 (p-value < 0.01). (4) Prediction evaluation, true positives of predicted coevolving pairs are evaluated using the *in vitro* and *in vivo* datasets (**Table S 6.1**). (5) Construction of Gag-protease coevolution networks. Gag-protease mutation patterns in the coevolution networks are classified into three categories: (a) true positives, predictions confirmed by *in vitro* or *in vivo* datasets; (b) unconfirmed positives, predictions that are not identified by either *in vitro* or *in vivo* datasets; (c) unpredicted positives, position pairs in *in vitro* or *in vivo* datasets that are not predicted. (6) Evaluation of coevolving Gag and protease positions in a six-step process. (a) The PDB dataset is used for analyzing the location of these positions in the HIV-1 protein structures. (b) The sequence and treatment data of 44 patients in the Leuven cohort is analyzed to identify Gag substitutions emerging during PI treatment. (c) The 137 drug binding positions of 50 Gag inhibitors are compared with the 29 PI-associated Gag positions in the Gag-protease coevolution networks. (d) HIV-1 Gag positions that interact with human proteins (**Table S 6.2**) are compared with the 29 PI-associated Gag positions. (e) Human CD4/CD8/antibody epitope positions in the HIV-1 Gag (**Table S 6.3**) are compared with the 29 PI-associated Gag positions. (f) The amino acid diversity of 420 drug-susceptible and 94 drug-resistant genomic sequences is compared to identify regions in the HIV-1 full-length genome associated with drug selective pressure.

6.4 Results

The Gag-protease coevolution networks

We modeled the Gag-protease coevolution networks using the full-length Gag and protease sequence datasets. Thirty-one Gag-protease coevolving pairs were predicted by a combination of four sequence-based methods (CMPro [41], NCPS [42], PhyCMAP [43], RCW [44]). Thirteen predictions were confirmed as true positives by the independent *in vitro* and/or *in vivo* datasets (A431+I54, A431+L10, A431+A71, A431+V82, A431+L33, L449+L10, A431+K20, A431+M46, A431+L90, L449+G16, A431+M36, I437+I54, L449+I54). For the other 18 coevolving pairs of Gag and protease positions, each harbored at least one Gag-protease mutation pattern (e.g.

A431V+I54V), whose prevalence differed significantly between the full-gagpro-resistant and the full-gagpro-susceptible datasets (p-value<0.01, **Table S 6.7**). **Figure 6.2** maps Gag-protease coevolving pairs, protein secondary structures, dN/dS, intra-subtype sequence diversity, human CD4/CD8/antibody T cell epitopes, Gag drug binding positions and Gag positions involved in the HIV-human protein interactions.

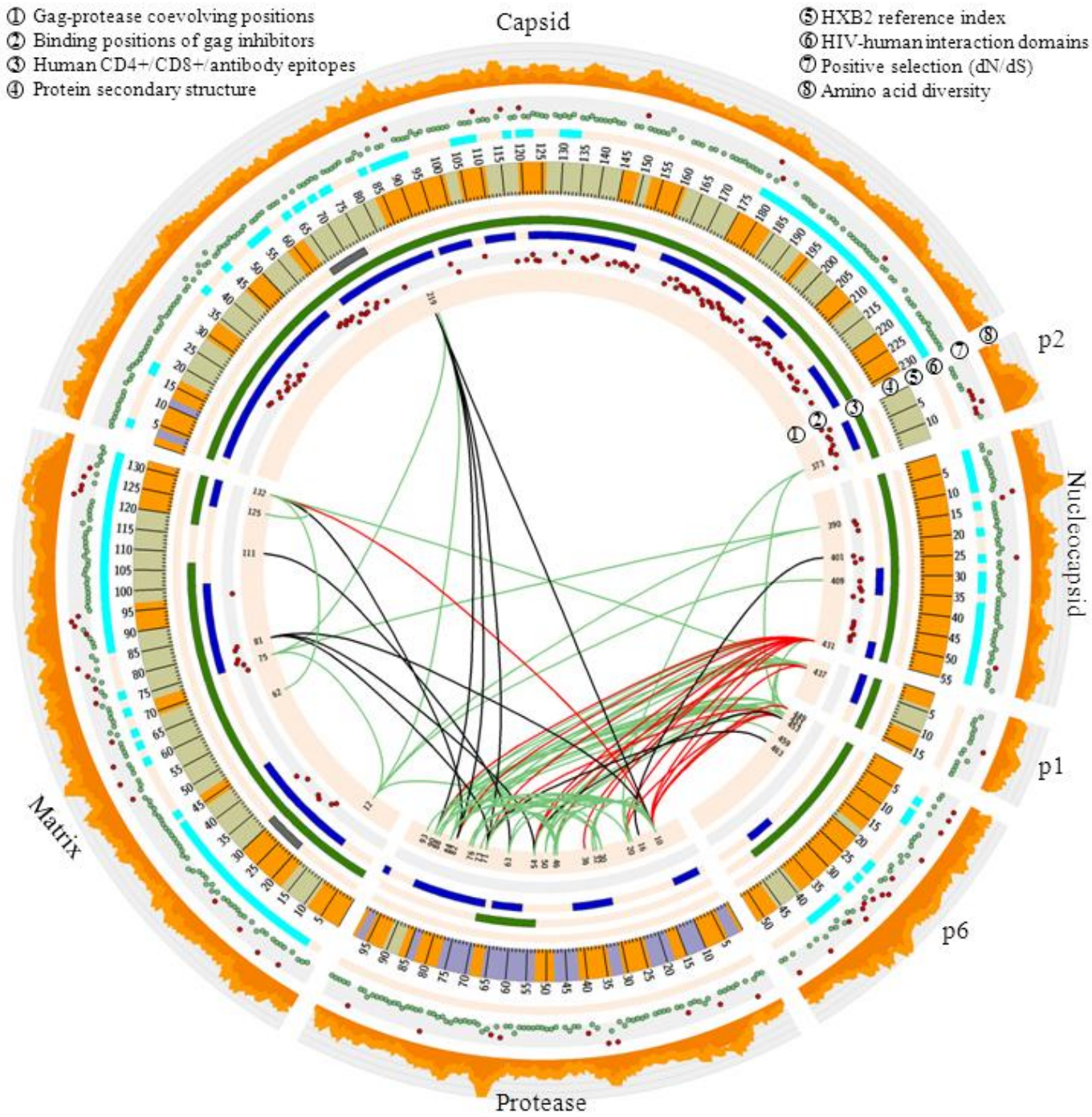


Figure 6.2: HIV-1 subtype B Gag-protease coevolution networks. Eight layers indicate: Layer 1: Gag-protease coevolving positions. Red lines indicate true positive

predictions confirmed by *in vitro* or *in vivo* datasets. Black lines indicate unconfirmed predictions. Green lines indicate position pairs in the *in vitro* or *in vivo* datasets that are not predicted. The index of Gag and protease positions are also annotated.
Layer 2: Drug binding positions of Gag experimental inhibitors (red dots).
Layer 3: Three sub-layers from outside to inside denote human antibody (grey), CD4+ (green) and CD8+ (blue) T cell epitopes, respectively.
Layer 4: Protein secondary structures. Colored regions indicate alpha-helix (blue), beta-strand (grey) and random-coil (orange) secondary structures.
Layer 5: Index of HIV-1 protein positions based on the HXB2 reference.
Layer 6: HIV-human protein interaction domains (skyblue).
Layer 7: dN/dS. Red dots indicate positively selected positions (dN/dS>1, p-values < 0.01); others are colored green.
Layer 8: Amino acid diversity of Gag and protease positions.
Visualization software: Circos (http://circos.ca/).

Gag-protease coevolution networks included 29 Gag positions associated with 30 protease positions. Among these 59 positions, 61.0% were highly variable (intra-subtype sequence diversity > 0.1) and 11.9% were under positive selection (dN/dS > 1, p-value < 0.01). Moreover, 22.0%, 33.9% and 44.1% of these 59 positions were located within alpha-helix, beta-strand and random-coil secondary structures, respectively (**Table S 6.5**). The top-ranked Gag-protease coevolving pairs included 4 Gag cleavage positions (A431, I437, L449, P453) that were previously confirmed by at least five studies, as well as 7 protease positions (L10, G16, I54, L63, A71, V82, L90) containing PI resistance mutations in four HIV-1 drug resistance interpretation algorithms (**Table S 6.4**). Moreover, a significant association was not found between PI-associated Gag positions and human antibody, CD8+ and CD4+ T cell epitopes (p-value>0.1), suggesting that immune selective pressure was unlikely to impact on the Gag-protease coevolution.

Four non-cleavage site positions (S111, V218, T401, F463) were newly identified to coevolve with protease residues (**Table S 6.5**). (1) For S111, Gag-protease coevolving pair S111+L63 harbored two mutation patterns (S111C+L63A, S111C+L63A) significantly associated with genotypic PI resistance (p-value<0.005). (2) For V218, four Gag-protease coevolving residues (V218P+L10I, V218P+V82A, V218P+A71V, V218P+L90M) had a high prevalence in the PI-resistant sequences (>1%, **Table S 6.5**). (3) For I401, coevolving pattern I401T+G16E was found in 3.29% of PI-resistant sequences, but was completely absent in the PI-susceptible sequences (p-value<0.01). (4) For F463, F463L+A71V was significantly associated with genotypic PI resistance (p-value < 0.0001).

HIV-1 Gag and protease mutations associated with PI drug susceptibility

We collected *in vitro* and *in vivo* data documented in literature to show the PI drug susceptibility of Gag and protease mutation patterns (**Table S 6.1**). Most Gag mutations (88.6%, 39/44) were associated with PI drug susceptibility in the presence of protease mutations (fold change IC50 > 1 or EC50 > 1, **Figure 6.3**).

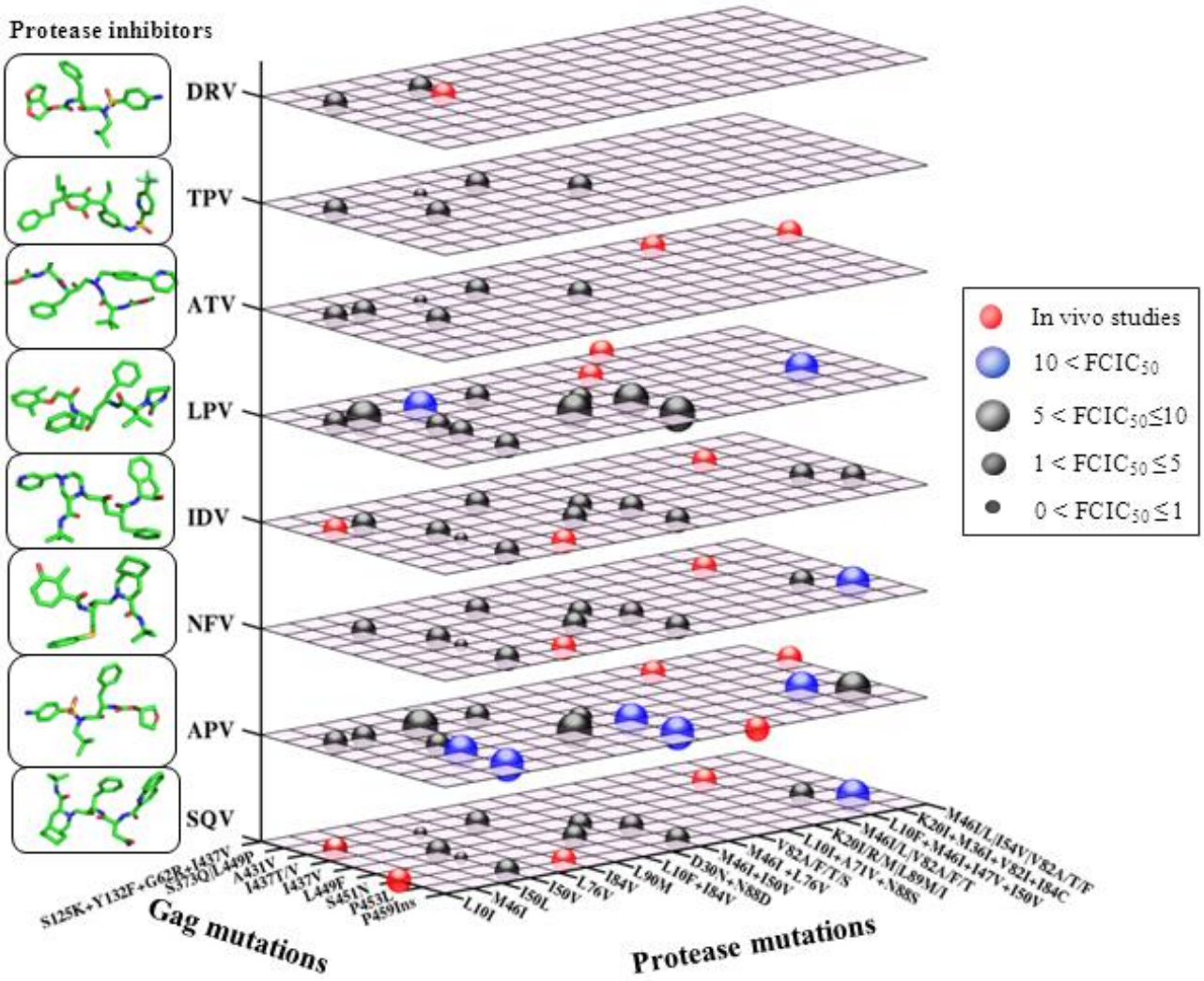


Figure 6.3: PI susceptibility associated with HIV-1 Gag-protease mutation patterns. X-, y- and z-axes represent Gag mutation patterns, protease mutation patterns and the level of PI susceptibility, respectively. The PI drug susceptibility of these Gag and protease mutation patterns are either reported in *in vitro* studies (fold change IC_{50} > 3 or EC_{50} > 3) or *in vivo* studies (**Table S 6.1**). Eight PIs extracted from crystallized structures of protease-inhibitor complexes are shown next to the z-axis. Red spheres indicate *in vivo* data and spheres in other colors indicate *in vitro* data. The legend indicates the degrees of PI susceptibility conferred by *in vitro* studies. PDB codes used for structural visualization: 1SDU (PI: IDV), 3EKX (NFV), 2NNP (SQV), 2Q5K (LPV), 3OXX (ATV), 2O4N (TPV), 3OXW (DRV), 3OXV (APV).

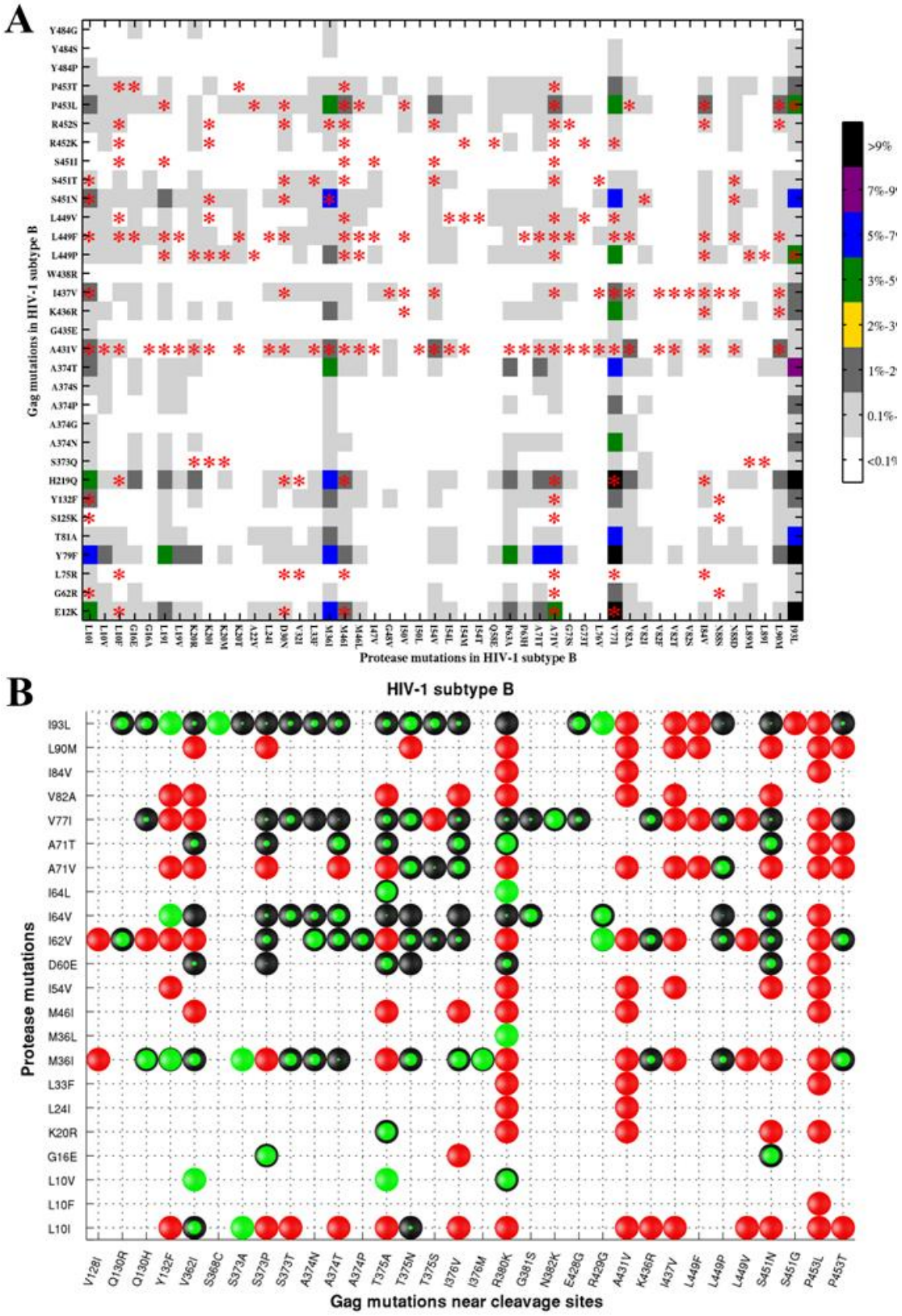


Figure 6.4: (A) Prevalence of Gag-protease mutation pairs in the HIV-1 subtype B Gag-protease sequence dataset. Gag and protease mutations are listed vertically and horizontally, respectively. Prevalence of Gag-protease mutations is colored according to the legend. Red asterisks indicate Gag-protease mutation pairs documented in *in vitro* or *in vivo* datasets (**Table S 6.1**).

(B) Association between CSMs and protease mutations in HIV-1 subtype B. Red circles indicate that the prevalence of CSM-protease mutation patterns is significantly differed between the PI-resistant and PI-susceptible sequence datasets (p-value<0.01). Other Gag-protease mutation patterns are indicated in black (p-values > 0.01) with scaled concentric green circles (1≤OR≤5) or full green circles (OR ≥5).

Of the 244 Gag-protease mutation pairs documented by *in vitro* (fold change IC50 > 1 or EC50 > 1) or *in vivo* dataset (mutations selected under PI selective pressure), 27.2% had a prevalence of less than 0.1% (**Figure 6.4**A). Of the 105 CSM-protease mutation pairs whose prevalence differed significantly between the PI-resistant and PI-susceptible sequence datasets (**Figure 6.4**B, p-value < 0.01), 44 were also documented in the *in vitro* and *in vivo* datasets. Particularly, seven Gag mutations (V128I, R380K, A431V, I437V, L449F/V, P453L) were strongly associated with protease mutations L24I, L33F, M46I, I54V, V82A, I84V and L90M (p-value < 0.01).

Gag substitutions emerging under PI selective pressure

To assess our Gag-protease coevolution networks, we used our Leuven patient cohort to report Gag amino acid substitutions that emerged after PI exposure. A longitudinal dataset of 44 patients who received PI-based regimens for at least 4 weeks was analyzed. Viruses in 6 patients developed Gag substitutions in the presence or the absence of protease substitutions (**Figure 6.5**). HIV-1 subtype B viruses from 3 PI-treated patients (ID: 133, 268, 290) developed Gag substitutions in the presence of protease substitutions. For patient 133 who received PI-based regimen for 65 weeks, two matrix substitutions V46T and Q63K emerged along with protease substitution T74S. For patient 268 who received PI-based regimen for 408 weeks, Gag substitution T470A emerged in the presence of protease substitution D30N. For patient 290, two Gag substitutions A431V/P453L and three protease substitutions L10F/I54V/V82A reverted back to wild type residues when a LPV/r-based regimen was interrupted for 21 weeks. On the other hand, viruses in 3 PI-treated patients (ID: 289, 314, 681) developed Gag substitutions in the absence of protease substitutions (**Figure 6.5**). For patient 289 who received PI-based regimen for 21 weeks, Gag substitutions N382H/S473A emerged while L483M reverted to wild type L483. Gag substitutions I401T/H421P/D425E/E482D were found in patient 314 who received a LPV/r-based regimen for 52 weeks. Gag substitution P478T was found in patient 681 who received PI-based regimens for 172 weeks.

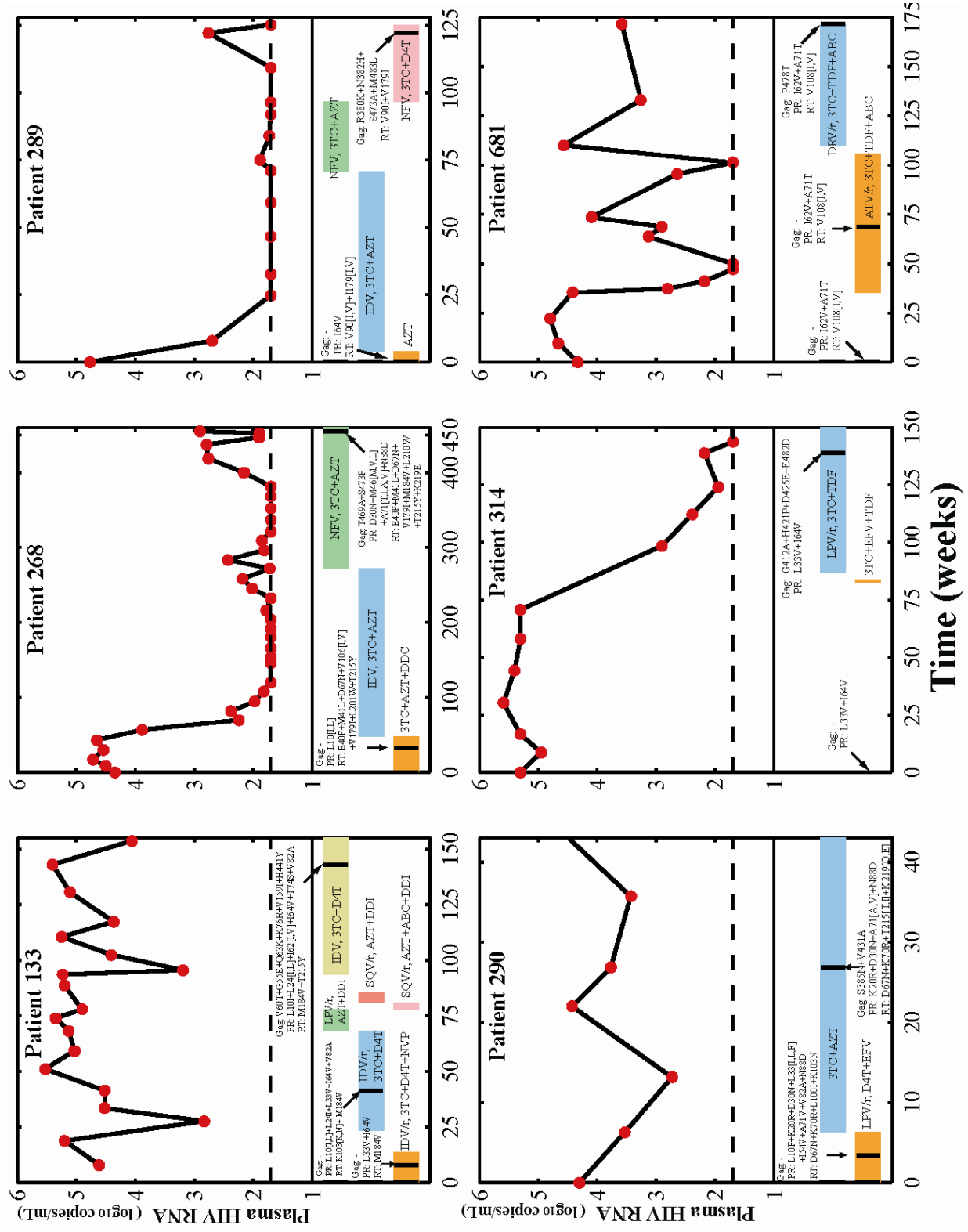


Figure 6.5: HIV-1 subtype B Gag and protease substitutions emerging during PI treatment. Each of six subplots shows emerging Gag-protease substitutions, viral loads and therapy records in a patient. The x- and y-axes display treatment time (weeks) and plasma viral load ($\log_{10}$ copies/mL), respectively. A colored box encompasses treatment time for each therapy regimen. Wild-type residues and substitutions in HIV-1 Gag and protease are annotated at the sampling date of

sequences. For convenience, the slash symbol "+" indicates multiple substitutions (e.g. I401T/H421P indicates the presence of both I401T and H421P). The plus symbol "+" separates Gag and protease substitutions (e.g. Q63K+T74S indicates Gag substitution Q63K and protease substitution T74S).

Gag C terminus and drug-targeted regions associate with drug selective pressure

The Gag C terminus (positions: 362-500) contains 18 (62.1%) of 29 PI-associated Gag positions, 24 of the 31 predicted Gag-protease coevolving pairs and 8 of the 10 Gag substitutions identified in our Leuven cohort. A significant difference of amino acid diversity between the 420 drug-susceptible and 94 drug-resistant genomic sequences was found at Gag cleavage sites (positions: 128-137, 359-368, 373-382, 428-437, 444-453) (p-value=0.0052), but not in the full-length Gag (p-value = 0.17). As shown in **Figure 6.6**, the amino acid diversity of Gag C terminus, protease and RT (position: 362-819) was significant higher in the drug-resistant (mean: 9.71%) than in the drug-susceptible genomes (mean: 7.39%, two-sample t-test, p-value = 1.2×10^{-62}). Amino acid diversity of other genomic regions did not differ significantly (13.57% vs. 13.83%, p-value = 0.104).

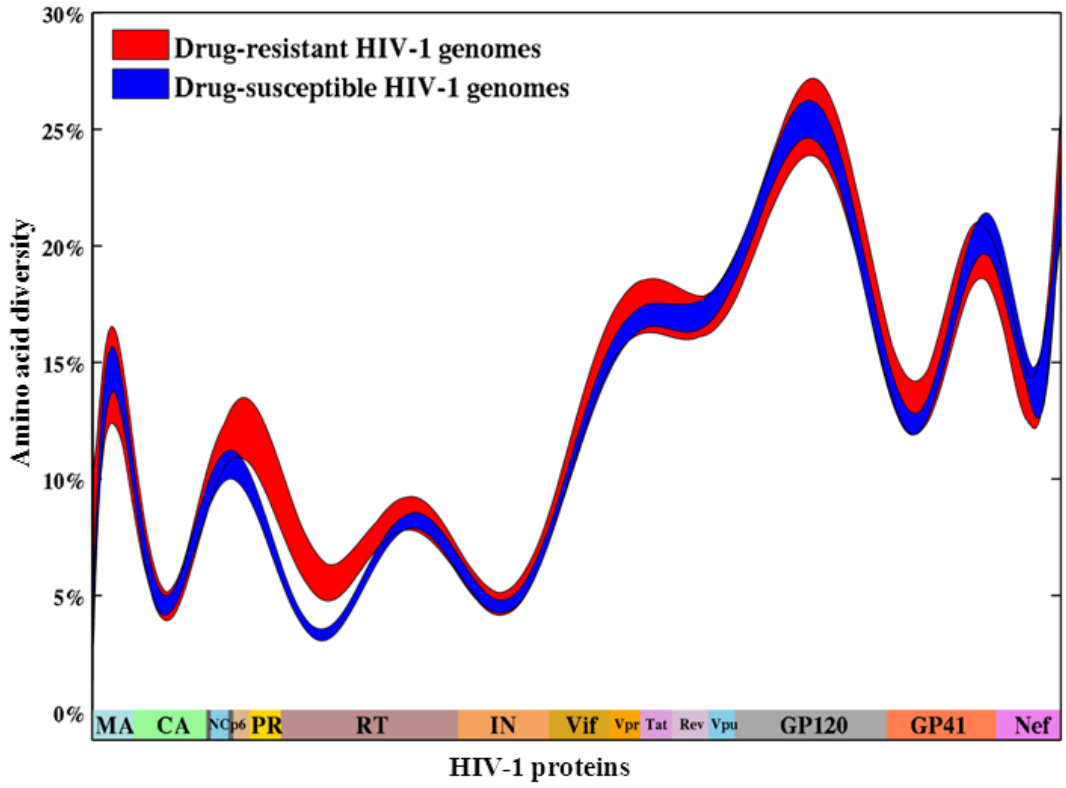


Figure 6.6: Distribution plots of amino acid diversity in the HIV-1 subtype B genome. Using bootstrap resampling with 1000 replicates, amino acid diversity of PI-susceptible and PI-resistant genomic sequences is calculated and colored in blue and

red, respectively. Protein names are indicated beneath the plots and concatenated protein regions are mapped based on the HXB2 reference.

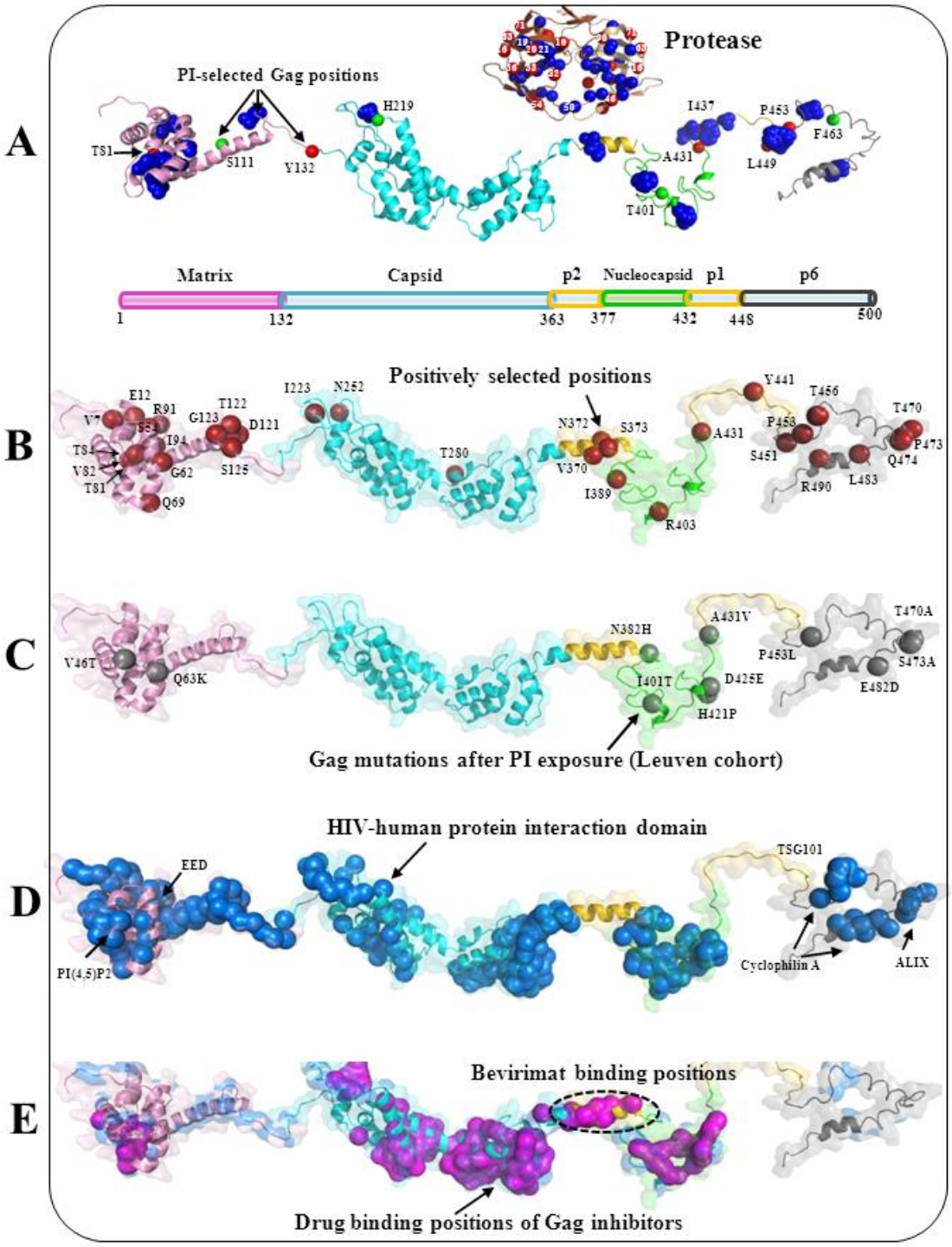


Figure 6.7: Structure representations of HIV-1 Gag and protease proteins.
(A) Mapping of PI-associated positions at the protein structures of HIV-1 protease (top) and Gag (bottom). The Cα atoms of amino acid positions are colored as red spheres (true positive predictions), green spheres (unconfirmed predictions) and blue spheres (positions from *in vitro* or *in vivo* datasets that are not predicted).
(B) Gag positions under positive selection (dN/dS>1, p-value<0.01).

(C) Gag substitutions identified after PI exposure in our Leuven cohort. Eleven Gag substitutions are annotated and the corresponding positions are colored in grey.
(D) HIV-human protein interaction domains. Skyblue spheres indicate amino acid positions in Gag that interact with human proteins (see summary in **Table S 6.2**).
(E) Drug binding positions of Gag experimental inhibitors. Red spheres map 137 drug binding positions in the full-length Gag. Three Gag inhibitors (PF-3450074, CAI and CAA) are annotated near their drug binding positions (see others in [25]).
PDB codes of Gag and protease proteins: 1HIW (matrix), 3NTE (capsid), 1U57 (p2), 2M3Z (nucleocapsid), 2C55 (p6), 1TW7 (protease). PDB codes of Gag inhibitors: 2BUO, 2L6E, 2XDE, 4E91, 4E92, 2JPR and 4INB. Visualization software: PyMOL V1.5 (http://www.pymol.org/).

Half PI-associated Gag positions interact with human proteins

Using the dataset of HIV-human protein interaction positions documented in literature (**Table S 6.1**), we identified 13 of 29 PI-associated Gag positions in the Gag-protease coevolution networks that interacted with human proteins (E12, L75, R76, S111, S125, Y132, V218, H219, T401, R409, P453, P459, P484) (**Figure 6.7**D). For instance, position E12 interacts with embryonic ectoderm development protein, position R76 interacts with PI(4,5)P2 and 5 positions (S125, Y132, H219, P453, P459) interact with Cyclophilin A (**Figure 6.7**). Of the 50 Gag cleavage sites, 6 positions (V128-Y132, Q135) interact with Cyclophilin A and 5 positions (A364-S368) interact with APOBEC3G (**Figure 6.7**).

Most PI-associated Gag positions do not interact with Gag experimental inhibitors

We compared 29 PI-associated Gag positions with 137 drug binding positions of 50 Gag experimental inhibitors. As illustrated in **Figure 6.7**E, only 6 PI-associated positions (L75, R76, T81, V390, T401, R409) interacted with two matrix inhibitors (Compound7 [45], TD1 [46]) and two nucleocapsid inhibitors (CAA [47], Compound6 [48]). Note that only the CA-p2 site in five Gag cleavage sites is targeted by Bevirimat and its analogs [49, 50], while the other cleavage sites that interact with human proteins are not the drug targets (**Figure 6.7**E).

6.5 Discussion and conclusions

To our knowledge, this study presents the first model of HIV-1 Gag-protease coevolution networks under PI selective pressure. Our findings demonstrate that Gag positions mostly in the cleavage sites and C terminus coevolve with protease drug resistance positions. Amino acid substitutions at coevolving Gag and protease positions can be selected during PI-based regimens (**Figure 6.5**). Moreover, the selective pressure of anti-HIV inhibitors could impact the sequence variability of Gag C terminus and drug-target regions in the subtype B genome. Surprisingly, many PI-associated Gag positions interact with human proteins, while only a few bind with Gag experimental inhibitors. Overall, our study contributes to the understanding of HIV-1 Gag-protease coevolution under PI selective pressure, shedding light on the impact of protein-protein coevolution in HIV-1 drug resistance.

HIV-1 Gag-protease coevolution provides a novel substrate-based mechanism for virus to escape PI selective pressure [5]. Documented by previous studies, a few Gag mutations (e.g. CSMs) can increase or decrease PI susceptibility, mostly in the presence of protease resistance mutations [6]. This process depends on the mutation combinations and the type of protease inhibitors (**Figure 6.3**). We mapped the coevolving Gag positions to protein secondary structures and found that 23 (79.3%) of the 29 PI-associated Gag positions were located in the flexible random-coil structures (**Table S 6.5**), supporting our hypothesis that the flexible Gag structure may play a crucial role in the HIV-1 Gag-protease coevolution. Our sequence-based method estimated 13 (41.9%) of 31 predictions as true positives, though some documented Gag-protease patterns were not predicted. Three factors may limit our predictions: (1) our sequence datasets did not contain all Gag-protease patterns reported by *in vitro* or *in vivo* studies (**Table S 6.1**). (2) Sequence-based methods mostly predict pairs of coevolving residues and underestimate high-dimensional associations between multiple residues [7]. (3) Sequence-based methods may fail to detect significant coevolving residues at highly conserved positions [51].

We showed that HIV-1 subtype B Gag and protease substitutions emerged under PI selective pressure using the longitudinal sequence data from our Leuven patient cohort. We found that 11 Gag substitutions in 13.6% (6/44) of subtype B infected patients who received PI-based regimens for at least 4 weeks. Of these 11 Gag substitutions, A431V and P453L were known and nine substitutions (N382H, I401T,

H421P, D425E, A431V, P453L, T470A, S473A, E482D) were located within the Gag C terminus. Our genome-wide sequence analysis further demonstrated that the amino acid diversity of Gag C-terminal domains and drug-target proteins was associated with genotypic PI resistance. This supported our findings in the Gag-protease coevolution networks that most PI-associated Gag mutations were located in the Gag C terminus. It also confirms the knowledge that drug resistance mutations in HIV-1 drug-target regions can be selected under antiviral treatments, causing a higher genetic diversity in viral sequences [52, 53]. While we did not detect significant variations of sequence diversity outside the Gag and Pol regions, it is possible that other regions may play a role in PI drug resistance due to genome-wide interactions. For instance, the interaction between matrix and GP41 may affect PI drug resistance – a hypothesis supported by the observation in two HIV-infected patients that some mutations at the GP41 cytoplasmic tail may confer PI drug resistance [54]. Our study could not perform a comparison analysis because the information of subtype and mutation was lacking in [54]. Despite these, our Gag-protease coevolution networks included two matrix mutations (L12E, Q62R), which involved with Env incorporation to nascent viral particles [55]. Interestingly, both matrix mutations can affect PI drug susceptibility *in vitro* [56, 57], while their roles in matrix-GP41 coevolution require further investigations.

We found that half (13/29) PI-associated Gag positions were located within the HIV-human protein interaction domains (**Figure 6.7**). Previous studies showed that Gag mutants have lower affinities to interact with human proteins compared to wild-type strains [58, 59]. It is possible that Gag positions exposed on the protein surface are accessible for protease binding and human protein interactions, since many human proteins interact with Gag proteins to stabilize viral proteins [1]. Future studies need to address the impact of Gag-protease coevolution on the HIV-human protein coevolution and vice versa.

More than 50 Gag experimental inhibitors have been published to date, and some have been under clinical trials [25]. Our coevolution analysis revealed that most Gag drug binding positions (95.6%, 131/137) were not coupled with protease residues, except a few detected at drug binding positions of Bevirimat (H358, L363, A364, A366, Q369, A370, T371) [49] (**Figure 6.7**E). Despite this, drug binding positions of

the promising capsid inhibitors (e.g. PF-3450074[60]), which target the conserved interaction interfaces of capsid N and C terminus, are unlikely to coevolve with protease residues (**Figure 6.7**). This supports the hypothesis that conserved N- and C-terminal domains in capsid can be potential drug targets for novel inhibitors [25].

Predicted Gag-protease coevolving pairs do not necessarily imply protein interaction positions; neither do they directly predict PI treatment failure. The limited number of PI-treated patients restricted our analysis from identifying all PI-associated Gag mutations, and the impact of individual Gag mutations on PI resistance and viral load response. Future studies need to investigate the role of the identified Gag substitutions in PI treatment failure. While we attempted to be as comprehensive as possible, additional *in vitro* and *in vivo* studies on Gag-protease mutations may have been reported, but major changes of known Gag-protease mutations are not expected. Since our study focused on the Gag-protease coevolution in the HIV-1 subtype B due to data availability, future studies are needed to characterize non-B subtypes. Besides the Gag-protease coevolution, other coevolution in the HIV-1 genome may have affected PI resistance. Future investigations of genome-wide analysis on HIV-1 drug resistance are still required.

6.6 Additional figures and tables

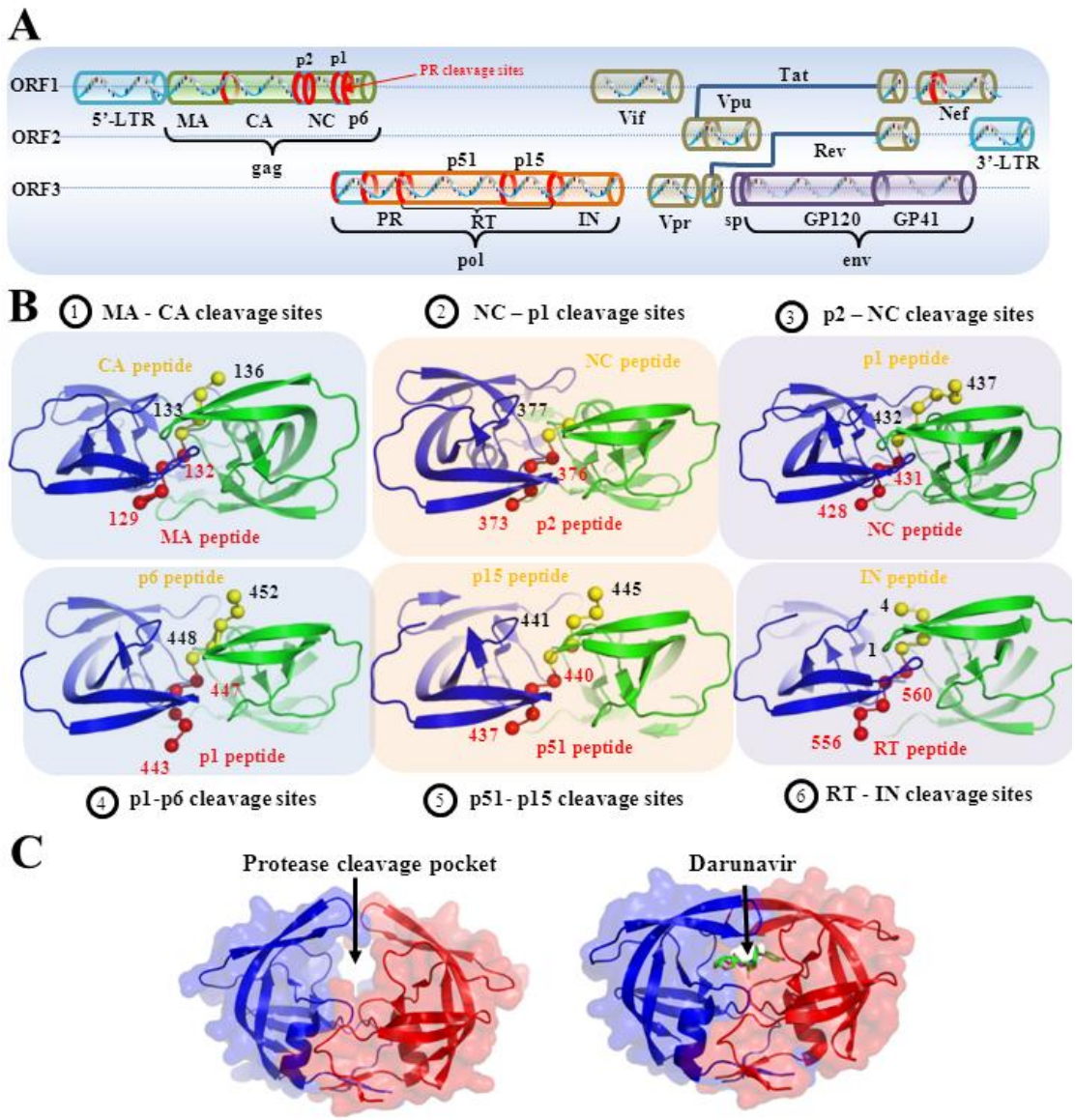


Figure S 6.1: Protease cleavage sites in the HIV-1 genome and protein structures of HIV-1 protease and its substrate peptides. (A) Schematic view of protease cleavage sites in the HIV-1 genome. HIV-1 full genome encodes 15 HIV-1 proteins and 3 space peptides in three open reading frames (ORFs). Protease cleavage sites are marked with red rings. (B) Protease structures crystalized with its substrate peptides. Subfigures from ① to ⑥ visualize the crystalized protease structures with substrate peptides derived from Gag cleavage sites (amino acid number: 10). Two units of HIV-1 protease dimers are colored blue and green, respectively. Upstream and downstream amino acids of Gag cleavage site are colored red and yellow, respectively. PDB codes used in the ① to ⑥ subfigures: 1KJ4, 1KJ7, 1TSQ, 1KJF, 1KJG and 1KJH. (C) Structure representations of wild type protease (left) and protease- darunavir complex (right). Visualization software: PyMOL V1.5.

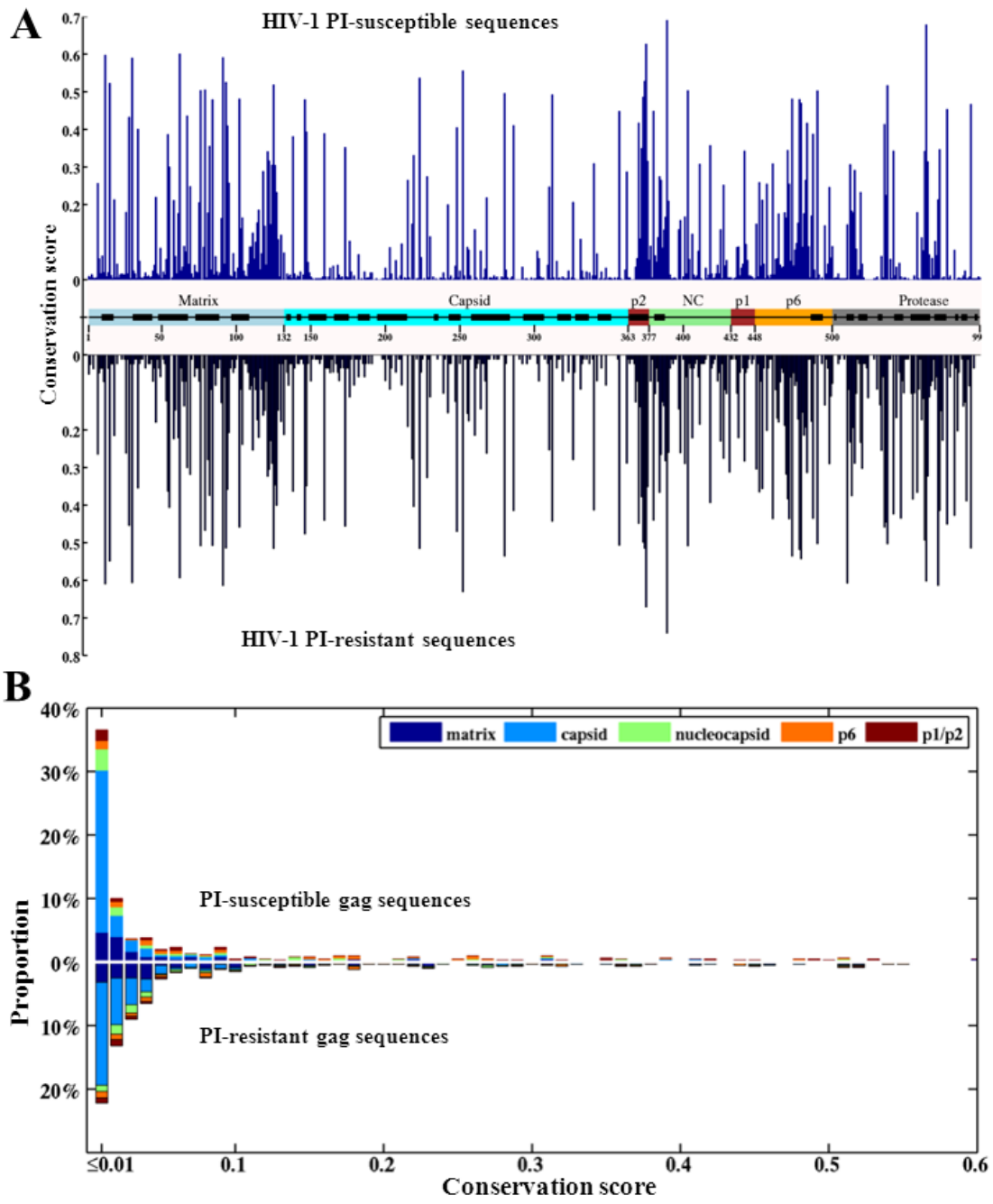


Figure S 6.2: Distribution of conservation scores in HIV-1 subtype B Gag and protease. (A) Conservation scores at individual Gag and protease proteins. The x- and y-axes represent Gag or protease HXB2 positions and conservation scores, respectively. For each Gag position, conservation scores calculated for PI-susceptible and PI-resistant sequences are shown symmetrically to the x-axis [25]. (B) Proportions of conservation scores in HIV-1 Gag proteins using PI-susceptible (top) and PI-resistant (bottom) sequences.

Table S 6.1: PI drug susceptibility of Gag and protease mutations measured using *in vitro* experiments

Gag mutations (1)	PR mutations (1)	RC%	Measure(2)	TPV	SQV	APV	DRV	NFV	IDV	LPV	RTV	ATV	Clone	Ref
R76K	None		FC EC50			~1.5		~1.2	~1	~0.9		~1.3	HXB2	[61]
Y79F	None		FC EC50			~1.2		~1	~0.8	~0.8		~1.1	HXB2	[61]
T81A	None		FC EC50			~2.5		~2.1	~1.6/1	~1.2		~2.9	HXB2	[61]
Y79F+T81A	None		FC EC50			~2.7		~3.7	~4	~2.3		~3.9	HXB2	[61]
R76K+T81A	None		FC EC50			~3.8		~3.6	~2.4	~2.1		~7	HXB2	[61]
R76K+Y79F	None		FC EC50			~2.3		~1.9	~1.7	~1.6		~2.2	HXB2	[61]
R76K+Y79F+T81A	None		FC EC50	~1.8(3)	~2.7	~5.7	~2	~7.5	~3.5	~3.5		~5.3	HXB2	[61]
L449F+P453T	L10F+G16E+K20T+A28S+M46I+A71V		FC IC50		2	8.7		5	55	15		1	HXB2	[62]
R452K	Q58E+A71V		FC IC50		1.1	0.4		1.0	0.9	1.0		1.4	HXB2	[62]
L449F	L10F	36	FC IC50		1.3	1.8		1.3	1.5	1.6	1.1		NL4-3	[63]
L449F	L10F+I84V	4	FC IC50		3.7	8.4		3.4	3.2	7.7	4.0		NL4-3	[63]
L449F	L10F+M46I+I50V	6	FC IC50		1.3	21		2.9	1.6	19	5.2		NL4-3	[63]
L449F	L10F+M46I+I47V+I50V	4	FC IC50		2.0	48		4.1	2.4	31	6.5		NL4-3	[63]
R452S	L10F+K20I+M36I+M46I+I54V+L63P+A71V+G73S+I84A	5.7	FC IC50		1000	43		196	178	50	400	700	Isolate A13	[64]
L449F	L10F+M46I+L63P+A71V+I84A	2.5	FC IC50		241	32		600	58	14	42		Isolate A4	[63, 64]
A431V+L449Q	L10V+M46I+L63A/P+A71V+I84A	18	FC IC50		256	28		600	53	9.0	54		Isolate A5	[64]
A431V	L10I+M46I+L63P+A71V+L76V+I84A	1.0	FC IC50		24	21		64	38	14	34		Isolate A6	[64]
A431V+S451I	L10F+L19I+M46I+I47V+I54V+L63P+A71V+I84A		FC IC50		1000	74		600	92	124	141		Isolate A7	[64]
A431V	L10V+K20I+M36I+M46I+A71V+G73S+L76V+I84A	6.4	FC IC50		265.5	252		600	400	94	86		Isolate A8	[64]
L449F	L10I+M46I+L63P+A71V+V77I+I84A	64	FC IC50		1000	33		600	138	16	167	52	Isolate A9	[64]
A431V	L10I+M46I+L63H+A71V+V77I+I84A	41	FC IC50		1000	32		600	73	10	60	38	Isolate A1	[64]
A431V+L449V+R452K	L10F+K20I+M46I+I54M+L63P+A71V+G73T+V77I+I84A	27	FC IC50		1000	400		600	400	59	400	700	Isolate A10	[64]
A431V+L449F	L10F+M46I+L63P+A71V+V77I+I84A	31	FC IC50		1000	47		600	73	14	73	63	Isolate A11	[64]
A431V+L449Q+S451T	L10I+L33F+M46I+I54V+L63P+A71V+L76V+I84A	16	FC IC50		1000	400		600	400	210	400	25	Isolate A12	[64]
R452S	L10F+K20I+M36I+M46I+I54V+L63P+A71V+G73S+I84A	5.7	FC IC50		1000	43		196	178	50	400	700	Isolate A13	[64]
S451N	K20I+M36I+V82I+I84C	11	FC IC50		14	5.4		59	1.3		8.6		Isolate C2	[64]
L449F	L10I+L19I+L24I+L63H+I84C	8.9	FC IC50		56	8.6		74	4.1	1.6	10		Isolate C3	[64]
A431V	L10I+G16A+M46I+L63P+L76V+I84C	13.0	FC IC50		24	64		144	63	47	22		Isolate C6	[64]
A431V+L449F	L10F+L19V+L24I+M46L+L63P+I84C	0.6	FC IC50		92	10		395	11	3.1	9.1		Isolate C4	[64]
A431V+L449F	L10I+L24I+M46L+L63P+A71T+G73S+V77I+I84C	1.2	FC IC50		1000	23		600	42	7.2	60		Isolate C7	[64]
A431V	M46I+L76V	10	FC IC50	0.7	0.3	6.6	1.4			10.6		0.4	HXB2	[65]
A431V	M46I		FC IC50	~3	~1.1	~3	~1.0			~3		~1.2	HXB2	[65]
A431V	L76V		FC IC50	~0.5	~0.5	~5.5	~1			~11.5		~0.5	HXB2	[65]
P453L	D30N+N88D		FC IC50					0.91					NL4-3	[66]
N165K	None		FC IC50	~3/1.4(4)	~2.8/2	4/3	~2/1.2	~6/3	~6/2	~4/2	~4.5/3	~3/2	Gag5(01AE)	[67]
E12K+L75R+H219Q+V390D+R409K	L10F+D30N+K45I+A71V		IC50 nM			0.64		11					NL4-3	[56]
E12K+L75R+H219Q+V390D+R409K	D30N+M46I+V77I		FCIC50			0.15		8					NL4-3	[56]
E12K+L75R+H219Q+V390D+R409K	None		IC50 nM		23	32		28	32		32		NL4-3	[56]
L75R+H219Q+V390D	L10F+V32I+M46I+I84V		FC p24			>1							NL4-3	[68]
K436E+I437T	None		FC IC50	~3.5	~3.9	~4.3		~5	~3.5	~4.7	~3.3	~3.3	HXB2	[5]
I437V	None		FC IC50	~2.7		~2.7				~5		~4.4	HXB2	[5]

Gag mutations (1)	PR mutations (1)	RC%	Measure(2)	TPV	SQV	APV	DRV	NFV	IDV	LPV	RTV	ATV	Clone	Ref
I437T	None		FC IC50	~3.8		~3.8				~5.5		~4.3	HXB2	[5]
A431V	L10F+K20T+L33F+M36I+M46I+I54V+L63P	~130	IC50 nM			~1250/400			~1200/400	~800/400			NL4-3	[69]
A431V	L10I+K20R+L33F+M46L+I54L+L63P+A71V+G73S+V82A+L90M	~70	IC50 nM			~210/50			~1000/150	~170/40			NL4-3	[69]
I437V	L10I+G48V+I54V+L63P+V77I+V82A	~24	IC50 nM			~350/220			~40/11	~130/70			NL4-3	[69]
L483P+K490R	L10I+I15V+E34Q+M36I+T37N+I54A+Q58E+V82A	42	IC50 nM						260/30		1650 /870		D1.10	[70]
I376V+ L483P+K490R	L10I+I15V+E34Q+M36I+T37N+I54A+Q58E+V82A	10	IC50 nM						110/30		60 /870		D1.10	[70]
I376V+E398V+L483P+K490R	L10I+I15V+E34Q+M36I+T37N+I54A+Q58E+V82A	14	IC50 nM						200/30		5150/870		D1.10	[70]
K494R	None	103	IC50 nM						10/30		9/870		D1.10	[70]
L449P+P453L	L19Ins+E21D+A22V+M46I/L63P+A71V+I84V+I93L	~30	FC IC50		46	17		>290	16	3.8	94	8.9	NL4-3	[71]
L449P+P453L	M46I+L63P+A71V+I84V+I93L	~100	FC IC50		3.3	3.0		10	1.3	1.8	8.5	2.6	NL4-3	[71]
	L19Ins+E21D+A22V		FC IC50		1.1	0.4		4.4	0.6	0.3	2.4	0.5	NL4-3	[71]
	M46I+I50V	6	FC IC50			10.8				6.4			HXB2	[72]
L449F	I50V	1	FC IC50		0.7	10.2		0.7	0.7	3.5	2.7		HXB2	[72]
L449F	M46I+I50V	16	FC IC50		1.0	15.2		3.8	1.8	7.8	6.8		HXB2	[72]
P453L	I50V	<1	FC IC50		1.0	10.8		1.3	1.3	4.8	3.7		HXB2	[72]
P453L	M46I+I50V	6	FC IC50		1.0	15.9		3.3	1.6	7.8	6.9		HXB2	[72]
K436E+I437T	None		FC EC50	~3.4						~4.6	~3.5		HXB2	[73]
G435R+K436E+I437T	None		FC EC50	~1.1						~1.2			HXB2	[73]
K436E+I437T+W438R	None		FC EC50	~1.3						~2.4			HXB2	[73]
A431V	None		FC EC50	~1.8						~2.6	~3.9		HXB2	[73]
I437V	None		FC EC50	~1.1						~1.9			HXB2	[73]
I437T	None		FC EC50	~1.1						~1.5			HXB2	[73]
A431V	L90M	602	FC IC50	10.31/4.59	79.98/26.4	44.1/17.2		114.55/29.7	78.48/16.52	108.87/25.56	200.53/49.64	77.76/25.62	NL4-3	[74]
A431V	N88D+L90M	48	FC IC50	11.0/4.5	471.0/125.0	26.0/18.4		451.0/230.0	62.9/17.3	48.7/20.1	139.0/53.1	217.9/46.1	NL4-3	[74]
A431V	N88D	138	FC IC50	5.1/1.9	185.0/30.2	14.1/5.1		293.0/109.0	27.3/5.8	25.5/5.6	58.2/13.4	106.7/17.2	NL4-3	[74]
A431V	D30N+N88D+L90M	103	FC IC50	11.7/3.8	572.2/135.8	28.3/21.3		555.1/305.9	70.5/17.4	45.5/15.5	121.4/38.3	258.2/52.1	NL4-3	[74]
A431V	D30N+N88D	122	FC IC50	1.5/5.0	27.3/196.8	4.7/14.2		117.2/320.7	5.2/26.8	4.1/21.9	9.0/44.5	16.9/112.4	NL4-3	[74]
A431V	I84V	1957	FC IC50	16.4/12.6	403.7/302.9	89.4/66.7		117.3/67.5	77.0/43.9	125.9/80.7	239.8/180.7	178.1/114.6	NL4-3	[74]
A431V	V82A	1775	FC IC50						41.4/47.9	90.4/102.2			NL4-3	[74]
A431V	I50L	58	FC IC50			5.5/2.0		9.9/3.5	3.9/1.3	4.8/0.9	9.0/1.9	123.4/42.8	NL4-3	[74]
A431V	I84V+L90M	1439	FC IC50	16.1/12.4	497.0/374.7	98.4/76.0		137.1/81.1	90.0/51.0	136.5/91.7	263.3/202.9	205.6/131.3	NL4-3	[74]
A431V	V82A+L90M	871	FC IC50	5.6/4.3					52.9/61.6		264.7/233.3		NL4-3	[74]
K436R	L90M	189	FC IC50	10.22/5.59	56.83/35.48	44.8/20.88		63.37/45.94	47.38/27.8	69.62/40.46	134.29/77.05	56.97/34.66	NL4-3	[74]
K436R	I84V+L90M	167	FC IC50								280.9/230.1	236.5/166.7	NL4-3	[74]
K436R	I84V	229	FC IC50								254.7/207.9		NL4-3	[74]
K436R	I50V	60	FC IC50		453.5/244.0			93.1/40.1	96.4/37.2	273.4/171.3			NL4-3	[74]
I437V	D30N+N88D	80	FC IC50	5.3/1.7	64.4/47.9	9.1/5.1		221.7/137.2	17.8/6.7	9.6/5.6	21.9/12.4	42.7/26.6	NL4-3	[74]
I437V	I50V	104	FC IC50	2.4/1.2	516.7/197.6	124.0/87.7		104.4/30.2	119.8/24.7	286.4/156.2	305.0/191.4	65.7/19.1	NL4-3	[74]
I437V	I84V+L90M	347	FC IC50	19.1/13.5	504.3/426.7	108.8/83.9		141.2/108.1	94.8/68.0	140.4/111.3	284.1/225.7	220.9/164.7	NL4-3	[74]
I437V	V82A	635	FC IC50	5.2/3.2	147.6/63.9	37.4/28.7		97.2/59.2	64.5/40.5	130.2/89.1	237.6/178.7	109.3/58.0	NL4-3	[74]
I437V	V82A+L90M	378	FC IC50	6.9/4.5	284.1/238.0	48.4/40.7		120.0/82.6	75.1/52.5	147.6/111.1	284.1/238.0	130.1/82.0	NL4-3	[74]
I437V	I84V	495	FC IC50	21.3/13.5	413.8/345.8	99.1/74.8		118.2/92.0	84.3/58.3	136.5/99.7	261.8/202.6	192.9/143.1	NL4-3	[74]

Gag mutations (1)	PR mutations (1)	RC%	Measure(2)	TPV	SQV	APV	DRV	NFV	IDV	LPV	RTV	ATV	Clone	Ref
L449F	D30N+N88D	296	FC IC50		39.3/72.0	4.4/8.4			5.9/11.0				NL4-3	[74]
L449F	V82A	321	FC IC50	3.4/4.4	69.0/160.9	29.1/37.1		59.8/119.1	40.0/75.5	91.0/129.0	182.6/243.4	62.7/99.7	NL4-3	[74]
L449F	I50V	181	FC IC50						30.8/51.5	147.4/200.6	163.3/241.5		NL4-3	[74]
L449F	V82A+L90M	179	FC IC50	4.8/6.1	95.0/233.5	40.7/52.1		82.0/154.9	50.8/93.4	111.1/157.5	240.1/311.6	85.0/129.5	NL4-3	[74]
L449F	L90M		FC IC50		66.24/33.22	38.49/21.45		81.39/44.12				69.05/33.06	NL4-3	[74]
L449F	N88D	320	FC IC50	76.19/39.72	76.19/39.72	8.26/4.96		164.59/119.52	11.89/6.51			40.1/22.72	NL4-3	[74]
R452S	I84V+L90M	517	FC IC50		546.8/408.3	96.6/85.9		143.9/105.4	92.0/67.3	138.1/110.6	255.4/229.2	194.8/167.0	NL4-3	[74]
R452S	I84V	649	FC IC50		466.5/329.2	92.3/75.6			81.9/57.9	134.3/98.8	237.8/205.3	176.1/144.6	NL4-3	[74]
R452S	L90M		FC IC50		49.2/36.69								NL4-3	[74]
P453L	I84V	2948	FC IC50		372.6/276.0	81.4/68.3		101.7/78.3					NL4-3	[74]
P453L	I84V+L90M	2255	FC IC50		446.0/364.3			116.7/98.4					NL4-3	[74]
P453L	L90M	722	FC IC50	7.21/5.56		38.45/16.64			40.71/27.49	58.59/38.4	101.3/75.45	54.27/32.29	NL4-3	[74]
P453L	V82A	759	FC IC50	3.5/3.7	77.3/91.5	28.3/34.8		61.7/83.0	41.0/56.3	91.7/110.7	179.2/223.9	65.5/75.3	NL4-3	[74]
P453L	V82A+L90M	465	FC IC50		107.8/121.9					112.3/132.9	238.1/276.8		NL4-3	[74]

(1) AA mutations: the symbol "+" indicates multiple mutations presented simultaneously (e.g. "Y79F+T81A" indicates the presence of both mutation Y79F and T81A). (2) The susceptibility to TPV, SQV, APV, DRV, NFV, IDV, RTV and ATV is presented from column 5 to 13 (empty values indicate unavailable data). The measurements used for analyzing drug susceptibility are shown in column 4. (3) For publications in which raw data were not presented in tables or texts, the symbol "~" preceding the approximate values indicates data collected from figures (e.g. "~5" RC% indicates that the replication capacity is approximately 5%). (4) The symbol "/" separates HIV-1 Gag-protease mutant's drug susceptibility from that of the protease mutant (only specific mutations present in the protease but not in the Gag). For instance, 1250/400 IC50 (nM) indicates that the IC50 of the corresponding Gag-protease mutant is 1250nM, and the IC50 of wild type is 400nM.

Table S 6.1 (B): HIV-1 Gag and protease mutations identified in patient cohort studies

Gag mutations	Protease mutations	Treatment	Number of subtype B infected patients	Study cohort	Ref
A431V	NA (1)	DRV/r	124		[22]
A431V	M46I +L76V	LPV	15	AREVIR	[18]
A431V, I437V, L449F, R452S, P453L	NA		313 (160^{N})(2)	DHCS/Danish HIV Database	[75]
P453L	I84V	48 IDV,57 RTV,22 NFV, 42SQVr	102(16^{N})		[76]
A431V	M46I/L,V82A/F/T	48 IDV,57 RTV,22 NFV,42 SQV+RTV+NFV	102(16^{N})		[76]
A374G/T/N/P/S, Y484G/I/P/S	NA	LPV/r	56	MONARK	[77]
K436R, I437V, L449F, P453L	NA	LPV/r	56	MONARK	[77]
G435E,K436N,I437V,L449V, L449F,S440C	NA	≥ 1 PI	953(628^{N})	RESINA	[20]
I437A	V82A	≥ 1 PI	953(628^{N})	RESINA	[20]
I437V	G48V,I50V,I54A/V,V82A/T	≥ 1 PI	953(628^{N})	RESINA	[20]
P459Ins	V82A/F/T/S	APV	84	NARVAL (ANRS 088)	[78]
S451N	L10I	SQV, RTV	42		[79]
I437T/V	L76V	DRV	43		[80]
A431V	M46I	IDV+RTV/SQV	28		[81]
V128T/A/del,L449F,I437V	NA	FPV+ATV/r, SQV+ATV/r	29	2IP-ANRS 127	[82]
N382A	I15V	SQV+ATV+RTV	1	2IP-ANRS 127	[83]
S373P, A374del, T375N, R380K	NA		2	2IP-ANRS 127	[83]
A431V	M46L/I+I54V+ V82A	IDV	8		[84]
L449P,S451N,P453L	D30N+N88D	NFV	21 clones from a patient		[85]
A431V	L24I+M46I/L+I54V+V82A	HAART	500(275^{N})	RESINA	[86]
I437V	I54V+V82F/T/S	HAART	500(275^{N})	RESINA	[86]
L449V	I54M/L/S/T/A	HAART	500(275^{N})	RESINA	[86]
L449F+R452S+P453L	D30N+I84V	HAART	500(275^{N})	RESINA	[86]
P453L	V82A	HAART	500(275^{N})	RESINA	[86]
L449F,S451N/T	D30N+N88D		196(B>90%)	NARVAL	[19]
A431V	M46I/L,I54V,V82A/T/F	SQV+RTV	98		[87]
S373Q,L449P	K20I/R/M,L89M/I	SQV+RTV	98		[87]
S125K+Y132F+G62R+I437V	L10I+A71V+N88S	LPV+ATV	98		[57]
P453L	N88D	NFV	36(6^{N})		[88]

(1) NA: data is not available. (2) An upper case "N" in superscript indicates the number of PI-naïve patients enrolled in the study.

Table S 6.1 (C): Kinetic parameters for the protease-mediated hydrolysis of HIV-1 subtype B Gag and protease mutants.

Gag mutations	PR mutations	Experimental values
A431V	I84V	5.9/2.6/1.6/1.0
A431V	L90M	9.6/2.6/2.5/1.0
A431V	M46L	2.4/2.6/0.3/1.0
A431V	V82A	2.1/2.6/0.7/1.0
Q430R+A431V	I84V	7.9/20.1/1.6/1.0
Q430R+A431V	L90M	45.4/20.1/2.5/1.0
Q430R+A431V	M46L	11.7/20.1/0.3/1.0
Q430R+A431V	V82A	1.1/20.1/0.7/1.0
F448Y	M46L	2.6/1.3/0.6/0.8
F448Y	V82A	2.2/1.3/1.2/0.8
L449F	M46L	8.7/7.6/0.6/0.8
F448Y	V82A	2.2/1.3/1.2/0.8
L449F	L90M	22.3/7.6/2.1/0.8

Experimental values are specificity constant Kcat/Km for protease-mediated hydrolysis of HIV-1 HXB2 clones, derived from reference [89]. The symbol "/" separates experimental values obtained from clones with a/b/c/d: (a) Gag + PR mutations, (b) Gag mutations, (c) protease mutations, and (d) wild type (e.g. 5.9/2.6/1.6/1.0 for HIV-1 strains with Gag A431V and protease I84V mutations indicate Kcat/Km = 5.9 for A431V+I84V mutations, Kcat/Km=2.6 for A431V mutant, Kcat/Km=1.6 for I84V mutant, Kcat/Km=1.0 for wildtype HXB2).

Table S 6.1 (D): Literature reports of HIV-1 Gag mutations in the absence of protease mutations

Gag mutations	Literature summary	Ref
P455TAP	Duplication of P455TAP motif in p6, prevalent in NRTI-treated but not in naive patients, improves virion packaging resulting in more infectious variants [90, 91]. Gag P455TAP insertions may be related to sequence variations in HIV-1 envelope [92]. In a cohort of 547 drug-naïve and 213 HAART failure isolates, P455TAP accumulated for longer lengths and at higher frequencies in subtype C patients than in subtypes B and F1 [93].	[90, 91], [92], [94], [93]
M377I	M377I near p2/NC site can block downstream p2/NC cleavage, resulting in faster cleavage of the CA/p2 site and less infectious virions.	[95]
A431V, R429K	Treatment-associated CSMs in a drug-naïve cohort could lower the genetic barrier of first-line PI therapy. A431V in the absence of PR mutations was significantly associated with R429K. A431V and R452S were correlated with primary PI resistance in drug naive HIV-1.	[96]
Q369A/H,T371A	Q6A/H, V7A/M and T8A near p1 can confer different levels of resistance to maturation inhibitor bevirimat (BVM).	[97, 98]
Q369,V370,T371	Natural polymorphisms near the CA-p1 link (at positions 358, 363, 364, 366, and 369-371) caused BVM drug resistance in 389 subtype B patients.	[99]
Q7,L33,N37,L63,C67,H69 in p6*	Gag residues Q7, L33, N37, L63, C67 and H69 at non-active site near p6* may influence catalytic site conformation and regulate protease-substrate specificity.	[100]
Y132I	Gag mutation Y132I near the MA-CA link can abolish viral infectivity compared to the wild type NL4-3.	[101]
N394F/G	Gag N17F/G modulates NC cleavage during late viral infection by decreasing infectivity and exhibiting H9 replication defects.	[102]
L363M, A364M	Gag L363M,A364M cause resistance to protease inhibitor DSB (3-O-(3',3'-dimethylsuccinyl)-betulinic acid), which delays cleavage of CA-SP1 in Gag.	[103]
MA/CA, p1/p6	Insertions (TGNS, SQVN, AQQA, SRPE, APP, and/or PTAPP) near MA-CA and p1-p6 links can restore the enzymatic activity of mutant protease.	[104]
Cleavage site mutations	Using 30 and 25 full genome sequences in subtype B and C respectively, 7 of 12 CSMs were more diverse in subtype C than in subtype B [105]. Amino acid context near the cleavage sites in Gag and Gagpol are crucial to determine Gag cleavage rate [12]. Natural polymorphisms in NC and C-terminal cleavage sites can affect protease processing activity. Polymorphisms in sites 374 to 380 may delay the dissociations between PR and Gag [106]. CSMs influence the drug resistance and viral fitness in patients treated with PIs [107]. CSM emergence was not associated with virological rebound among patients treated with LPV/RTV in the OK04 clinical trial [108].	[105]
NC-p1	HIV-1 protease variants that evolved in reaction to treatment with RTV had 1.2-fold increased mean fitness, increased susceptibility to 2 NRTIs and SQV, and impaired replication capacity, correlated with reduced Gag NC-p1 processing.	[109]
Mutant Gag	Mutant Gags, derived from one patient with multi-PI drug-resistance, acted synergistically with mutant protease to reduce PI susceptibility while maintaining replication capacity. Non-CSM mutations in Gag alone can also reduce susceptibility to APV,ATV,DRV,IDV,LPV,NFV,SQV,TPV.	[110]
Gag and env mutations	Gag and env mutations are associated with PI antiviral treatment failure.	[111]
MA and CA mutations	Mutations in MA and parts of CA can reduce PI susceptibility and restore the loss of replication capacity of protease mutant.	[61, 110, 112]

Table S 6.2: Summary HIV-1 Gag-human protein interactions with identified binding domains

HIV-1 proteins	HIV-1 binding domains	Human host factors	Interaction function	Reference
Matrix	8-43[113],W16,W36[59]	ubiquitous calcium-sensing calmodulin (CaM, CAMI)	Gag intracellular Trafficking in cytoplasm	[113],[59],[114],[115]
	R4-L13,R20-E40[116]	embryonic ectoderm development (EED) protein	Transcriptional regulation within nucleus	[116],[117]

	S9-K28	IL-8 chemokine receptors CXCR1(IL-8RA)/CXCR2	Endothelial cells proangiogenic activation, monocyte migration	[118],[119],[120]
	K26-K28	elongation factor 1-alpha (EF1a)	Inhibition of translation	[121]
	121-132	Cyclophilin A(CypA)	Enhance CA- CypA interaction	[122]
	15-32	Chromosome maintenance region 1 (CRM1,exportin 1, XPO1)	MA nuclear export signal	[123]
	24-31, 110-114[124]	Importin α1 (karyopherin α2, Rch1/SRP1α/KPNA2)	PIC nuclear import to nucleus	[124]
	Y132	clathrin adaptor complex 2 AP-2, μ2 subunit (AP50)	Gag intracellular Trafficking	[125]
	5-8,13-16	TIP47 (tail-interacting protein of 47 kDa, perilipin 3, PLIN3)	Env packaging into virions	[126],[127]
	K26,K27	HO3 histidyl-tRNA synthetase (HARS2)	HO3 packaging into virions	[128]
	9,67,72,77[129]	mitogen-activated protein kinase (MAPK/ERK-2)	Phosphorylation of matrix during early uncoating, MAPK/ERK-2 incorporats to virions	[129],[130],[131]
	L41,F44,V46,I60,L64,L75	PS/PE/PC (phosphatidylserine, phosphatidylethanolamine,phosphatidylcholine)	Gag binding to membrane	[132],[133]
	113-122	Mab 3H7 antibody	PIC integration	[134]
	S111	protease kinase C (PKC)	matrix translocation to membrane	[135],[136]
	25-34,109-115	Heparin(HSPG analog)	Prevent p17 binding to chemochine receptor	[137]
	86-115	neutralizing monoclonal antibody (MAb 1575)		[138]
	S6[139],R22,K26,K27,W36,R76[140]	phosphatidylinositol-(4,5)-bisphosphate PI(4,5)P2	Target Gag to membrane rafts	[140],[141]
Capsid	P85-I91[142],T54,A92,R132[143],A92E,G94D[144],H219[145],P221,P222[146],N74D[147],P85-A88,A92,P93,G94	Cyclophilin A(CypA)	Viral core uncoating, incorporate into virions	[148],[142],[149],[143],[150],[151]
	N53,L56,N57,M66,Q67,K70,I73, N74,A105,T107,S109,Y130	CPSF6 (CFI_m)	CPSF6 binds CA in post-entry stages before Uncoating, nuclear import	[152],[153]
	G89[154],H87[155],P38[156], V83,G89,H120,P122,W117, Y130,W133[157]	TRIM5α	TRIM5α promotes capsid disassembly during viral uncoating in cytoplasm	[158],[154],[159],[155],[156],[160],[161],[157]
	E45,T54,N57,Q63,Q67,N74,A105[162]	Transportin3(TNPO3,TRN-SR2)	PIC nuclear import	[162], [163], [164]
	N74	Nucleoporin NUP98	PIC nuclear import	[165],[166],
	N57,Q67,K70,N74[167]	Nucleoporin NUP153	PIC nuclear import	[168] ,[167], [169]
	G89,P90,I91[170]	Nucleoporin NUP358 (RanBP2)	PIC nuclear import	[170],[171],[165],[172]
	S16,P17	peptidyl prolyl-isomerase PIN1	CA core uncoating in cytoplasm	[173]
	V3	clathrin adaptor complex 2 AP-2, μ2 subunit	Gag intracellular Trafficking	[125]
	V3	adapter protein complex 2 AP-2, α1 subunit	Nuclear translocation of viral DNA in cytoplasm or perinuclear region	[174]
	177-231	lysyl-tRNA synthetase LysRS	LysRS packed into virions	[175]
Nucleocapsid	K34,C49,N55	Moloney leukemia virus 10 (MOV10)	MOV10 packaging during virion budding	[176],[177]
	R3,R7,R10,K11,K14,K20,R26[58]	ALIX (AIP-1)	Recruit Gag to plasma membrane in viral budding	[58],[178],[179],[180],[181],[179]
	M1-K11[182],R29-K34[183]	APOBEC3G (A3G)	A3G incorporation to virions in viral budding	[182],[184],[185]
	Y36-P49	mRNA binding protein 1 (IMP1)	Impedes Gag assembly, keep immature virus on cellular membranes	[186]
	C15-C49	double-stranded RNA-binding protein Staufen 1 (Stau1)	Stau1 packed into virions, influences Gag multimerization	[187]
	K14,K20,R26,R29,K33,K34,K38, K41,K47	ATP-binding protein ABCE1 (HP68)	capsid assembly	[188]
	43-48	Topoisomerase I(TOP1)	Enhancing reverse transcription	[189]

	K14,K20,R26,R29,K32-K34,K38,K41,D48	Elongation factor 1-alpha (EF1a)	Inhibition of translation	[121]
	R3,R7,R10,K11,K14,K20,R26	Nedd4-like ubiquitin ligase, Nedd4-1	Viral release	[179]
p6	Y36-L41[190],E34,L35,P37,L41,R42[191]	ALIX (ALG-2 interacting host protein, AIP-1)	HIV-1 buds via the Alix driven pathway, ALIX incorporates into virions	[58],[192],[179],[180],[191]
	P7-P10,	Tumor susceptibility gene 101(TSG101)	Form viral budding machinery to bud from plasma membrane	[58], [193]
	K27	small ubiquitin-like modifier SUMO-1	ESCRT-III recruitment to viral budding	[194]
	K27	Ubc9	ESCRT-III recruitment to viral budding	[195], [194]
	K27	Daxx	ESCRT-III recruitment to viral budding	[194]
	T23	mitogen-activated protein kinase (MAPK/ERK-2)	P6 phosphorylation within HIV-1 virion	[196]
	L35-L38	Nedd4-like ubiquitin ligase, Nedd4-1	Viral release	[179]
	P5,P7,P10,P11,P24,P30,P37,P49	Cyclophilin A(CypA)	Catalyzes prolyl cis/trans interconversion of p6 Pro residues	[197]
	K27,K33	Ubiquitin	assembly and budding	[198]

Note that positions in HIV-1 binding domains are referred to HXB2 reference at individual Gag protein. Information on space peptide p1 and p2 is not available, for we could not find human factors bind to them.

Table S 6.3: Summary of human antibody, CD4+, CD8+ T cell epitopes in HIV-1 Gag

HIV-1 protein	Antibody epitope position	CD4+ epitope position	CD8+ epitope position
Matrix	20-31	1-107,118-132	11-44,74-101,124-132
Capsid	64-75	1-231	3-56,61-92,94-104,108-117,121-153,161-189,197-205,217-231
p2		1-14	1-10
Nucleocapsid		1-55	28-36,50-55
p1		1-16	1-10
p6		1-43	33-41
Protease		53-70	3-11,30-42,57-66,68-90,99

Table S 6.4: HIV-1 protease positions in HIV-1 genotypic drug resistance interpretation algorithms

Expert rules	Protease positions	Version	Year	Reference
IAS-USA	10,11,16,20,24,30,32,33,34,36,43,46,47,48,50,53,54,58,60,62,63,64,69,71,73,74,76,77,82,83,84,85,88,89,90,93	March,2013	2013	[199]
HIVdb	10,11,20,24,30,32,33,35,36,43,46,47,48,50,53,54,58,63,71,73,76,77,82,83,84,85,88,89,90,93	6.2.0	2012	[200]
Rega	10,11,20,24,30,32,33,35,36,43,46,47,48,50,53,54,58,62,63,64,71,73,74,76,77,82,84,85,88,89,90,93,95	9.1	2013	[201]

ANRS	10,15,20,24,30,32,33,36,46,47,48,50,53,54,62,63, 71,73,76,77,82,84,88,90	22	2012	[202]

Protease positions are listed in the HIV-1 genotypic drug-resistance interpretation algorithms, which can be found via the following links:
(a) IAS-USA: https://www.iasusa.org/content/hiv-drug-resistance-mutations
(b) HIVdb: http://hivdb.HIVdb.edu/DR/cgi-bin/rules_comments_hivdb.cgi?class=PI
(c) Rega: http://rega.kuleuven.be/cev/avd/software/rega-algorithm
(d) ANRS: http://www.hivfrenchresistance.org/.

Table S 6.5: Summary of HIV-1 Gag and protease positions reported by literature or our study

Gag protein	Position	Proportions of Gag mutations (1)	Closest Gag cleavage sites (2)	Intrasubtype Diversity(3)	dN/dS(4)	Secondary structure	Reference(5)
Matrix	12	E12K(24.65%)		0.6007	8.37(p<0.01)	Alpha-helix	[56]
Matrix	62	G62R(4.04%)		0.5183	5.27(p<0.01)	Alpha-helix	[57]
Matrix	75	L75R(0.32%)		0.2387	0.502(p=0.949)	Alpha-helix	[56]
Matrix	76	R76K(51.86%)		0.5193	2.11(p=0.0393)	Alpha-helix	[6]
Matrix	79	Y79F(40.70%)		0.4908	1.32(p=0.31)	Alpha-helix	[61]
Matrix	81	T81A(8.82%)		0.1929	6.52(p<0.01)	Alpha-helix	[61], our study
Matrix	111	S111C(5.22%), S111G(1.07%)		0.2518	0.971(p=0.631)	Alpha-helix	Our study
Matrix	125	S125K(1.81%)		0.5351	3.15(p<0.01)	Random-coil	[57]
Matrix	132	Y132F(5.01%)	MA-CA	0.1075	0.701(p=0.841)	Random-coil	[57], our study
Capsid	218	V218A(3.69%), V218P(4.20%)		0.2083	1.23(p=0.308)	Random-coil	our study
Capsid	219	H219Q(20.90%)		0.4244	3.32(p=0.0208)	Random-coil	[56] [68]
P2	370	V370A(14.34%),V370I(2.27%), V370M(4.99%)		0.486	2.72(p<0.01)	Random-coil	[99]
P2	373	S373Q(1.53%)	P2-NC	0.5305	4.44(p<0.01)	Random-coil	[83], [87]
P2	374	A374G(1.46%),A374N(6.02%), A374P(4.74%),A374S(2.64%),A 374T(13.86%),A374V(2.01%)	P2-NC	0.5507	1.41(p=0.109)	Random-coil	[83] [77]
P2	375	T375N(18.82%)	P2-NC	0.6015	1.62(p=0.0361)	Random-coil	[83]
NC	390	V390D(<0.1%),V390I(12.55%)		0.2828	1.56(p=0.232)	Random-coil	[56]
NC	401	T401L(7.24%),T401T(1.48%),T 401V(2.36%)		0.2763	2.99(p=0.221)	Random-coil	our study
NC	409	R409K(<0.1%)		0.079	0.0145(p=1)	Random-coil	[56]
NC	410	K410R(2.77%)		0.048	0.67(p=0.829)	Random-coil	our study
NC	431	A431V(2.11%)	NC-p1	0.0398	4.12(p<0.01)	Random-coil	[64] [65] [73] [74] [22] [18] [75] [76] [81] [86] [87], our study
P1	435	G435E(0.25%),G435R(<0.1%)|	NC-p1	0.0211	0.5(p=0.889)	Random-coil	[20] [73] [203]
P1	436	K436E(<0.1%),K436N(<0.1%), K436R(4.73%)	NC-p1	0.1248	0.0811(p=1)	Random-coil	[5] [73] [74] [77] [20]
P1	437	I437A(<0.1%),I437T(<0.1%),I4 37V(2.87%),I437L(2.28%)	NC-p1	0.1039	0.967(p=0.621)	Random-coil	[5] [73] [74] [75] [77] [20] [80] [82] [86], our study
P1	438	W438R(0.25%)		0.0119	0.066(p=0.997)	Random-coil	[73]
P1	440	S440C(<0.1%)		0.0375	0.47(p=0.931)	Random-coil	[20]
P6	449	L449F(1.36%),L449P(7.26%),L4 49Q(<0.1%),L449V(1.49%)	p1-p6	0.1984	0.551(p=0.979)	Random-coil	[63] [64] [71] [72] [74] [204] [75] [77] [20] [82] [85] [86] [19] [87], our study
P6	451	S451I(0.13%),S451N(14.60%),S 451T(0.38%),S451G(1.56%)	p1-p6	0.3258	Inf(p<0.01)	Random-coil	[79] [19]
P6	452	R452K(0.75%),R452S(0.79%)	p1-p6	0.1278	0.203(p=1)	Random-coil	[62] [64] [74] [86]
P6	453	P453L(7.79%),P453T(4.79%),P4 53S(1.68%)	p1-p6	0.245	Inf(p<0.01)	Random-coil	[62] [65] [71] [72] [74] [75] [76] [77] [85] [86], our study
P6	459	P459I(<0.1%)		0.1217	0.25(p=0.963)	Random-coil	[78]
P6	463	F463L(3.44%)		0.1101	3.98(p=0.121)	Random-coil	Our study
P6	484	Y484G(0.13%),Y484I(<0.1%),Y 484P(0.60%),Y484S(0.35%),Y4 84H(1.05%)		0.088	0.53(p=0.872)	Random-coil	[77]

Protease protein	Positions	Proportion of protease mutations	Drug resistance interpretation rules (6)	Intrasubtype diversity	dN/dS	Secondary structure	Reference
Protease	10	L10F(1.61%),L10I(10.47%),L10V(2.71%)	IAS,HIVdb,Rega,ANRS	0.1517	0.757(p=0.9)	Beta-strand	[62] [63] [64] [79],our study
Protease	16	G16A(0.63%),G16E(4.54%)	IAS	0.0536	0.167(p=1)	Random-coil	[62] [64], our study
Protease	19	L19I(7.60%),L19V(2.11%)		0.14	1.57(p=0.149)	Beta-strand	[71] [64] [71]
Protease	20	K20I(0.63%),K20M(1.07%),K20R(3.60%),K20T(0.66%)	IAS,HIVdb,Rega,ANRS	0.0624	0.587(p=0.97)	Beta-strand	[62] [64] [87], our study
Protease	21	E21D(<0.1%)		0.0022	0.332(p=0.95)	Beta-strand	[71]
Protease	22	A22V(<0.1%)		0.0009	0.111(p=1)	Beta-strand	[71]
Protease	24	L24I(0.95%)	IAS,HIVdb,Rega,ANRS	0.011	0.804(p=0.752)	Beta-strand	[64] [86]
Protease	28	A28S(<0.1%)		0.0013	Inf(p=0.667)	Random-coil	[62]
Protease	30	D30N(1.01%)	IAS,HIVdb,Rega,ANRS	0.0107	Inf(p=0.118)	Random-coil	[65] [74] [204] [85] [86] [19]
Protease	32	V32I(1.26%)	IAS,HIVdb,Rega,ANRS	0.0142	0.254(p=0.999)	Beta-strand	[137]
Protease	33	L33F(2.14%),L33I(1.07%),L33V(1.73%)	IAS,HIVdb,Rega,ANRS	0.0495	1.65(p=0.123)	Beta-strand	[71] [69], our study
Protease	36	M36I(16.65%),M536L(1.23%)	IAS,HIVdb,Rega,ANRS	0.1867	1.13(p=0.668)	Random-coil	[64] [69] [70],our study
Protease	37	S37D(10.00%),S537C(1.51%),S537S(16.11%),S537T(2.62%),S537H(1.42%)		0.356	7.23(p<0.01)	Random-coil	[70]
Protease	45	K45I(<0.1%),K545R(1.89%)		0.0199	0.446(p=0.929)	Beta-strand	[56],
Protease	46	M46I(3.78%),M46L(1.32%),M46P(<0.1%)	IAS,HIVdb,Rega,ANRS	0.0514	0.634(p=0.862)	Beta-strand	[62] [63] [64] [65] [68] [71] [72] [89] [18] [76] [81] [86] [87],our study
Protease	47	I47V(0.60%)	IAS,HIVdb,Rega,ANRS	0.0091	Inf(p=0.594)	Beta-strand	[63]
Protease	48	G48V(0.54%)	IAS,HIVdb,Rega,ANRS	0.0091	0.401(p=0.985)	Beta-strand	[20]
Protease	50	I50L(0.16%),I50V(0.16%)	IAS,HIVdb,Rega,ANRS	0.0038	Inf(p=0.709)	Random-coil	[20, 63] [72] [74]
Protease	54	I54A(<0.1%),I54L(0.60%),I54M(0.19%),I54S(<0.1%),I54T(0.16%),I54V(3.50%)	IAS,HIVdb,Rega,ANRS	0.0454	2.12(p=0.0721)	Beta-strand	[20, 64] [70] [84] [86] [87],our study
Protease	58	Q58E(1.17%)	IAS,HIVdb,Rega	0.0117	0.499(p=0.96)	Beta-strand	[62] [70]
Protease	63	L63A(5.42%),L63H(1.80%),L63P(55.41%)	IAS,HIVdb,Rega,ANRS	0.4185	1.04(p=0.497)	Beta-strand	[64] [71] ,our study
Protease	71	A71T(8.34%),A71V(8.53%)	IAS,HIVdb,Rega,ANRS	0.1721	2.99(p<0.01)	Beta-strand	[62] [64] [71],our study
Protease	73	G73S(1.17%),G73T(0.25%)	IAS,HIVdb,Rega,ANRS	0.0174	2.3(p=0.136)	Beta-strand	[64]
Protease	76	L76V(0.57%)	IAS,HIVdb,Rega,ANRS	0.0063	0.417(p=0.952)	Beta-strand	[65] [18] [80]
Protease	77	V77I(33.79%)	IAS,HIVdb,Rega,ANRS	0.3391	6.01(p<0.01)	Beta-strand	[64] [56] [69]
Protease	82	V82A(3.41%),V82F(0.28%),V82I(1.64%),V82S(0.13%),V82T(0.25%)	IAS,HIVdb,Rega,ANRS	0.0584	4.09(p<0.01)	Random-coil	[62] [74] [89] [76] [20] [78] [86] [87],our study
Protease	84	I84A(<0.1%),I84C(<0.1%),I84V(2.34%)	IAS,HIVdb,Rega,ANRS	0.0259	Inf(p=0.31)	Beta-strand	[63] [64] [68] [71] [74, 86] [89] [76]
Protease	88	N88D(0.98%),N88S(0.25%)	IAS,HIVdb,Rega,ANRS	0.013	0.331(p=0.998)	Alpha-helix	[65] [74] [204] [85] [19]
Protease	89	L89I(0.16%),L89M(0.98%)	IAS,HIVdb,Rega	0.02	0.552(p=0.954)	Alpha-helix	[87]
Protease	90	L90M(5.49%)	IAS,HIVdb,Rega,ANRS	0.0562	2.07(p=0.0307)	Alpha-helix	[74] [89] [204],our study
Protease	93	I93L(35.42%)	IAS,HIVdb,Rega	0.3586	4.83(p<0.01)	Alpha-helix	[71]

(1) Proportions of mutations at the corresponding Gag or protease positions in our HIV-1 Gag-protease sequence dataset (the most prevalent amino acids are considered as wild type amino acids, others are mutations). (2) The closest Gag cleavage sites are indicated if Gag positions are less than 5 amino acids away from the cleavage sites. (3) Intra-subtype sequence diversity (see Methods). (4) dN/dS, the ratio of non-synonymous and synonymous rates at a residue position (see Methods). If dS=0, then dN/dS is infinite, denoted as "Inf". (5) Reference, the list of publication that reported the Gag or protease positions, the results from our study are indicated by "our study". (6) HIV-1 drug resistance interpretation algorithms that reported protease drug resistance mutations.

Table S 6.6: Summary of Gag, protease and RT substitutions in 6 patients of the Leuven cohort.

Patient ID	Sampling day	Gag substitution	Protease variants and PI resistance mutations	RT variants and RTI resistance mutations
133	2000-03-06		L33V+I64V	M184V
	2000-10-30		L10[I,L]+L24I+L33V +I64V+V82A	K103[K,N]+ M184V
	2002-10-21	V60T+G55E+Q63K+ K76R+V159I+H441Y	L10I+L24[I,L]+I62[I,V] +I64V+T74S+V82A	M184V+T215Y
268	1996-07-10		L10[I,L]+L63[Q,H,P]	E40F+M41L+D67N+V106[I,V] +V179I+L201W+T215Y
	2005-05-18	T469A+S473P	L10[I,L]+D30N+M46[M,V,L] A71[T,I,A,V]+N88D	E40F+M41L+D67N+V179I+ M184V+L210W+T215Y+K219E
289	1997-09-24		I64V	V90[I,V] +I179[I,V]
	2000-02-09	R380K+N382H+ S473A+M483L		V90I+V179I
290	2006-03-07		L10F+K20R+D30N+ L33[I,L,F]+I54V +A71V+V82A+N88D	D67N+K70R+L100I+K103N
	2006-09-21	S385N+V431A	K20R+D30N+ A71[A,V]+N88D	D67N+K70R+T215[T,I] +K219[Q,E]
314	2005-07-18		L33V+I64V	K166R
	2008-03-31	G412A+H421P+ D425E+E482D	L33V+I64V	K166[K,R]
681	2009-08-19		I62V+A71T	V108[I,V]
	2010-12-20		I62V+A71T	V108[I,V]
	2012-12-20	P478T	I62V+A71T	V108[I,V]

For the first sequence, PR and RT drug resistance mutations, detected by in the drug resistance interpretation algorithms HIVdb v7.0 [26] and/or Rega V9.1 [27], are colored red. For the subsequent sequences, the amino acid changes regarding to the first sequence are displayed (emergence and disappearance of amino acids). Ambiguous nucleotide letters are translated into amino acids, which are indicated by brackets. Gag substitutions and PI mutations are mapped in **Figure 6.5** according to the sampling time.

Table S 6.7: Coevolving residue pairs in the HIV-1 Gag-protease coevolution networks

Position pair	Coevolution score	Mutation pattern	Proportion*	P-value	Odds-ratio	Confirmed by literature
New Gag-protease coevolving pairs predicted by our sequence analysis						
81+10	0.144	T81A+L10I	3.28%/0.31%	0.0036	10.7059	
		T81A+L10F	2.46%/0	0.0003	91.05	
		T81A+L10V	1.64%/0.47%	0.0676	3.5041	
81+71	0.138	T81A+A71V	3.31%/0.31%	0.0036	10.7966	
81+54	0.129	T81A+I54V	2.46%/0	0.0003	91.05	
		T81A+I54L	4.10%/0	<0.0001	154.322	
		T81L+I54V	1.64%/0	0.0026	60.1983	
111+63	0.148	S111C+L63A	4.10%/0.32%	0.0015	13.3475	

		S111C+L63S	1.64%/0	0.0025	60.1983	
132+82	0.125	Y132F+V82A	4.96%/0	<0.0001	188.3793	
218+10	0.181	V218P+L10I	4.92%/0	<0.0001	186.7692	
218+82	0.171	V218P+V82A	4.13%/0	<0.0001	155.641	
218+90	0.154	V218P+L90M	3.28%/0	<0.0001	122.4202	
		V218A+L90M	1.64%/0	0.0024	60.1983	
218+71	0.129	V218P+A71V	4.13%/0.30%	0.0012	14.188	
401+16	0.166	I401T+G16E	3.29%/0	<0.0001	123.0405	
453+54	0.155	P453L+I54V	7.35%/0	<0.0001	288.0678	
		P453L+I54L	1.84%/0	<0.0001	67.984	
		P453T+I54V	1.31%/0	0.0002	48.3024	
463+71	0.153	F463L+A71V	3.40%/0.17%	<0.0001	20.6784	
Gag-protease coevolving pairs identified by both our sequence analysis and literature results						
132+10	0.128	Y132F+L10I	5.74%/0	<0.0001	219.7759	
431+54	0.184	A431V+I54V	7.06%/0	<0.0001	274.8679	
		A431V+I54L	2.35%/0	0.0001	87.2335	
431+10	0.174	A431V+L10I	8.24%/0	<0.0001	324.7643	[64]
		A431V+L10F	2.35%/0	0.0001	87.2335	[64]
		A431V+L10V	1.18%/0	0.0036	43.1006	[64]
431+71	0.161	A431V+A71V	5.33%/0	<0.0001	203.5901	
		A431V+A71T	1.78%/0	0.0006	65.4251	
		A431V+A71I	1.18%/0	0.0036	43.3571	
431+82	0.149	A431V+V82A	7.69%/0	<0.0001	301.5669	[74],[76],[84], [86], [87]
		A431V+V82T	1.18%/0	0.0036	43.3571	
431+33	0.146	A431V+L33F	6.47%/0	<0.0001	250.3875	[62], [63]
431+20	0.144	A431V+K20R	4.12%/0	<0.0001	155.4512	
431+46	0.141	A431V+M46I	5.92%/0	<0.0001	227.625	[64],[65],[18],[76], [81],[84],[86],[87]
		A431V+M46L	2.96%/0	<0.0001	110.3636	[76],[89],[84],[86],[87]
431+90	0.138	A431V+L90M	8.07%/0	<0.0001	317.7584	[74]
431+36	0.129	A431V+M36I	8.38%/0	<0.0001	331.0909	
437+10	0.187	I437V+L10I	5.36%/0.26%	<0.0001	21.9375	
		I437V+L10F	1.79%/0	0.0006	65.8193	
437+54	0.127	I437V+I54V	4.17%/0	<0.0001	157.3704	
449+10	0.145	L449V+L10I	3.46%/0	<0.0001	130.0714	
		L449F+L10I	1.86%/0	<0.0001	68.9027	
		L449F+L10F	1.06%/0	0.0008	39.0563	[62], [63]
453+10	0.153	P453L+L10I	5.48%/0.58%	<0.0001	9.9866	
		P453L+L10F	4.96%/0	<0.0001	189.5836	
		P453L+L10V	1.83%/0	<0.0001	67.6233	
		P453A+L10I	1.04%/0	0.0009	38.3368	
		P453T+L10I	1.57%/0.29%	0.0045	5.496	
453+20	0.14	P453L+K20R	3.70%/0.14%	<0.0001	26.6384	[87]

		P453L+K20T	1.59%/0	<0.0001	58.5845	
453+71	0.13	P453L+A71V	8.71%/0.43%	<0.0001	21.8732	
		P453L+A71T	3.43%/0.43%	<0.0001	8.1471	
		P453L+A71I	1.06%/0	0.0008	38.7447	
		P453A+A71V	1.32%/0	0.0002	48.56	
		P453T+A71V	1.58%/0.29%	0.0044	5.5428	
		P453T+A71T	1.32%/0.22%	0.0066	6.1467	
Gag-protease coevolving pairs identified by *in vitro* and *in vivo* studies						
12+30	0.066	E12K+D30N	2.46%/0	0.0003	91.05	
373+20	0.154	S373P+K20R	2.10%/0	0.0004	77.4894	
		S373P+K20I	1.40%/0	0.0029	51.2958	
		S373P+K20T	1.40%/0	0.0029	51.2958	
431+76	0.084	A431V+L76V	1.76%/0	0.0006	65.0357	
431+88	0.058	A431V+N88D	1.24%/0	0.0034	45.525	
431+84	0.11	A431V+I84V	6.47%/0	<0.0001	250.3875	[64], [74]
436+90	0.036	K436R+L90M	1.24%/0	0.0034	45.525	
436+84	0.044	K436R+I84V	1.18%/0	0.0036	43.1006	
437+82	0.12	I437V+V82A	4.19%/0	<0.0001	158.3478	
437+84	0.076	I437V+I84V	1.79%/0	0.0006	65.8193	
449+82	0.121	L449F+V82A	1.07%/0	0.0008	39.2668	
		L449V+V82A	1.87%/0	<0.0001	69.2772	
449+90	0.091	L449F+L90M	4.36%/0	<0.0001	165.5455	
		L449P+L90M	1.63%/0	<0.0001	60.3646	
		L449V+L90M	3.00%/0	<0.0001	112.2185	
449+88	0.084	L449F+N88D	1.64%/0	<0.0001	60.5319	[74]
449+54	0.125	L449F+I54V	1.60%/0	<0.0001	59.2195	
		L449V+I54V	1.87%/0	<0.0001	69.2772	
451+10	0.14	S451N+L10I	4.77%/0.67%	<0.0001	7.3833	
		S451N+L10F	1.33%/0	0.0002	48.8204	
		S451T+L10I	1.33%/0	0.0002	48.8204	
452+84	0.056	R452S+I84V	2.62%/0	<0.0001	97.6408	[64]
452+90	0.049	R452S+L90M	2.67%/0	<0.0001	99.7808	[74]
453+50	0.051	P453L+I50V	1.05%/0	0.0009	38.438	
453+84	0.093	P453L+I84V	9.69%/0	<0.0001	389.4624	[71], [74], [76], [86]
		P453A+I84V	1.05%/0	0.0009	38.438	
453+90	0.089	P453L+L90M	11.50%/0	<0.0001	471.7048	[74]
		P453A+L90M	1.87%/0	<0.0001	69.2772	
		P453T+L90M	2.41%/0	<0.0001	89.5574	
453+82	0.119	P453L+V82A	2.62%/0	<0.0001	97.9032	[74], [86]
459+82	0.046	P459S+V82A	0.50%/0	<0.0001	0	[78]

*: Proportions of Gag-protease amino acid patterns were calculated using the PI-susceptible and PI-resistant Gag-protease sequence datasets.

Table S 6.8: The list of software used in this study

Software	Software function used in our study	Software availability	Input	Ref
Circos	Network visualization	http://circos.ca/		[205]
COMET	Subtype classification	http://comet.retrovirology.lu/	MSA	
HyPhy v2.1.0	Positive selection pressure	http://hyphy.org/w/index.php/Main_Page	MSA,Tree	[37]
Hypermut V2.0	Hypermutation sequence test	http://www.hiv.lanl.gov/content/sequence/HYPERMUT/hypermut.html	MSA	[206]
MUSCLE	Sequence alignment	http://www.drive5.com/muscle/	MSA	[207]
MultipleTest	Bergmann-Hommel's tests	http://sci2s.ugr.es/keel/multipleTest.zip		[208]
Matlab 2012a	Statistical tests	http://www.mathworks.com/products/matlab/		
PSIPRED V3.3	Secondary structure prediction	http://bioinf.cs.ucl.ac.uk/psipred/	MSA	[32]
PDBREPORT	PDB quality control	http://swift.cmbi.ru.nl/gv/pdbreport/	PDB	[31]
ProtTest3	AA substitution test	http://darwin.uvigo.es/software/prottest3/prottest3.html	MSA	[209]
PyMOL V1.5	Protein visualization	http://www.pymol.org/	PDB	[210]
PreRec V1.04	Precision-recall curve	http://www.mathworks.com/matlabcentral/fileexchange/29250		[211]
PSIPRED	Secondary structure prediction using MSA	http://bioinf.cs.ucl.ac.uk/psipred/	MSA	[32]
2Struc	Secondary structure prediction using PDB	http://2struc.cryst.bbk.ac.uk/twostruc	PDB	[212]
RAxML V7.0.4	ML phylogenetic tree	http://www.exelixis-lab.org/software.html	MSA	[213]
Rega subtype tool	Subtype classification	http://bioafrica.mrc.ac.za:8080/rega-genotype-3.0.2/hiv/typingtool	MSA	[24]
Seaview V4.3.2	Sequence editor	http://pbil.univ-lyon1.fr/software/seaview.html	MSA	[23]

6.7 References

1. Waheed AA, Freed EO. HIV type 1 Gag as a target for antiviral therapy. *AIDS Res Hum Retroviruses* 2012,**28**:54-75.
2. Ozen A, Haliloglu T, Schiffer CA. Dynamics of preferential substrate recognition in HIV-1 protease: redefining the substrate envelope. *J Mol Biol* 2011,**410**:726-744.
3. Ganser-Pornillos BK, Yeager M, Sundquist WI. The structural biology of HIV assembly. *Curr Opin Struct Biol* 2008,**18**:203-217.
4. Sadiq SK, Noe F, De Fabritiis G. Kinetic characterization of the critical step in HIV-1 protease maturation. *Proc Natl Acad Sci U S A* 2012,**109**:20449-20454.
5. Nijhuis M, van Maarseveen NM, Lastere S, Schipper P, Coakley E, Glass B, *et al.* A novel substrate-based HIV-1 protease inhibitor drug resistance mechanism. *PLoS Med* 2007,**4**:e36.
6. Fun A, Wensing AM, Verheyen J, Nijhuis M. Human Immunodeficiency Virus Gag and protease: partners in resistance. *Retrovirology* 2012,**9**:63.
7. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. *Nat Rev Genet* 2013,**14**:249-261.
8. Lovell SC, Robertson DL. An integrated view of molecular coevolution in protein-protein interactions. *Mol Biol Evol* 2010,**27**:2567-2575.
9. Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C. Transient protein-protein interactions: structural, functional, and network properties. *Structure* 2010,**18**:1233-1243.
10. Mintseris J, Weng Z. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci U S A* 2005,**102**:10930-10935.
11. Cimarelli A, Darlix J-L. HIV-1 Reverse Transcription. In: *Human Retroviruses*: Springer; 2014. pp. 55-70.
12. Lee SK, Potempa M, Kolli M, Ozen A, Schiffer CA, Swanstrom R. Context surrounding processing sites is crucial in determining cleavage rate of a subset of processing sites in HIV-1

Gag and Gag-Pro-Pol polyprotein precursors by viral protease. *J Biol Chem* 2012,**287**:13279-13290.
13. Liu Z, Wang Y, Brunzelle J, Kovari IA, Kovari LC. Nine crystal structures determine the substrate envelope of the MDR HIV-1 protease. *Protein J* 2011,**30**:173-183.
14. Chaudhury S, Gray JJ. Identification of structural mechanisms of HIV-1 protease specificity using computational peptide docking: implications for drug resistance. *Structure* 2009,**17**:1636-1648.
15. Tie Y, Boross PI, Wang YF, Gaddis L, Liu F, Chen X, *et al.* Molecular basis for substrate recognition and drug resistance from 1.1 to 1.6 angstroms resolution crystal structures of HIV-1 protease mutants with substrate analogs. *FEBS J* 2005,**272**:5265-5277.
16. Prabu-Jeyabalan M, Nalivaika EA, King NM, Schiffer CA. Structural basis for coevolution of a human immunodeficiency virus type 1 nucleocapsid-p1 cleavage site with a V82A drug-resistant mutation in viral protease. *J Virol* 2004,**78**:12446-12454.
17. Kalinina OV, Oberwinkler H, Glass B, Krausslich HG, Russell RB, Briggs JA. Computational identification of novel amino-acid interactions in HIV Gag via correlated evolution. *PLoS One* 2012,**7**:e42468.
18. Knops E, Kemper I, Schulter E, Pfister H, Kaiser R, Verheyen J. The evolution of protease mutation 76V is associated with protease mutation 46I and gag mutation 431V. *AIDS* 2010,**24**:779-781.
19. Kolli M, Lastere S, Schiffer CA. Co-evolution of nelfinavir-resistant HIV-1 protease and the p1-p6 substrate. *Virology* 2006,**347**:405-409.
20. Knops E, Brakier-Gingras L, Schulter E, Pfister H, Kaiser R, Verheyen J. Mutational patterns in the frameshift-regulating site of HIV-1 selected by protease inhibitors. *Med Microbiol Immunol* 2012,**201**:213-218.
21. Knops E, Daumer M, Awerkiew S, Kartashev V, Schulter E, Kutsev S, *et al.* Evolution of protease inhibitor resistance in the gag and pol genes of HIV subtype G isolates. *J Antimicrob Chemother* 2010,**65**:1472-1476.
22. Larrouy L, Lambert-Niclot S, Charpentier C, Fourati S, Visseaux B, Soulie C, *et al.* Positive impact of HIV-1 gag cleavage site mutations on the virological response to darunavir boosted with ritonavir. *Antimicrob Agents Chemother* 2011,**55**:1754-1757.
23. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 2010,**27**:221-224.
24. Pineda-Pena AC, Faria NR, Imbrechts S, Libin P, Abecasis AB, Deforche K, *et al.* Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: Performance evaluation of the new REGA version 3 and seven other tools. *Infect Genet Evol* 2013.
25. Li G, Verheyen J, Rhee SY, Voet A, Vandamme AM, Theys K. Functional conservation of HIV-1 gag: implications for rational drug design. *Retrovirology* 2013,**10**:126.
26. Liu TF, Shafer RW. Web resources for HIV type 1 genotypic-resistance test interpretation. *Clinical infectious diseases* 2006,**42**:1608-1618.
27. Van Laethem K, De Luca A, Antinori A, Cingolani A, Perna CF, Vandamme AM. A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients. *Antivir Ther* 2002,**7**:123-129.
28. Maes B, Schrooten Y, Snoeck J, Derdelinckx I, Van Ranst M, Vandamme AM, *et al.* Performance of ViroSeq HIV-1 Genotyping System in routine practice at a Belgian clinical laboratory. *J Virol Methods* 2004,**119**:45-49.
29. Van Laethem K, Schrooten Y, Dedecker S, Van Heeswijck L, Deforche K, Van Wijngaerden E, *et al.* A genotypic assay for the amplification and sequencing of gag and protease from diverse human immunodeficiency virus type 1 group M subtypes. *J Virol Methods* 2006,**132**:181-186.
30. Libin P, Beheydt G, Deforche K, Imbrechts S, Ferreira F, Van Laethem K, *et al.* RegaDB: community-driven data management and analysis for infectious diseases. *Bioinformatics* 2013,**29**:1477-1480.
31. Hooft RW, Vriend G, Sander C, Abola EE. Errors in protein structures. *Nature* 1996,**381**:272.
32. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. *Bioinformatics* 2000,**16**:404-405.
33. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999,**292**:195-202.
34. Fu W, Sanders-Beer BE, Katz KS, Maglott DR, Pruitt KD, Ptak RG. Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Res* 2009,**37**:D417-422.

35. Llano A, Frahm N, Brander C. How to optimally define optimal cytotoxic T lymphocyte epitopes in HIV infection. *HIV molecular immunology* 2009,**2009**:3-24.
36. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 2010,**5**:e9490.
37. Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 2005,**21**:676-679.
38. Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002,**64**:479-498.
39. Agresti A. *An introduction to categorical data analysis*: Wiley-Interscience; 2007.
40. Szumilas M. Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry* 2010,**19**:227.
41. Di Lena P, Nagata K, Baldi P. Deep architectures for protein contact map prediction. *Bioinformatics* 2012,**28**:2449-2457.
42. Lee BC, Kim D. A new method for revealing correlated mutations under the structural and functional constraints in proteins. *Bioinformatics* 2009,**25**:2506-2513.
43. Wang Z, Xu J. Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics* 2013,**29**:i266-273.
44. Gouveia-Oliveira R, Pedersen AG. Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. *Algorithms Mol Biol* 2007,**2**:12.
45. Zentner I, Sierra LJ, Fraser AK, Maciunas L, Mankowski MK, Vinnik A, *et al.* Identification of a small-molecule inhibitor of HIV-1 assembly that targets the phosphatidylinositol (4,5)-bisphosphate binding site of the HIV-1 matrix protein. *ChemMedChem* 2013,**8**:426-432.
46. Alfadhli A, McNett H, Eccles J, Tsagli S, Noviello C, Sloan R, *et al.* Analysis of small molecule ligands targeting the HIV-1 matrix protein-RNA binding site. *J Biol Chem* 2013,**288**:666-676.
47. Goudreau N, Hucke O, Faucher AM, Grand-Maitre C, Lepage O, Bonneau PR, *et al.* Discovery and Structural Characterization of a New Inhibitor Series of HIV-1 Nucleocapsid Function: NMR Solution Structure Determination of a Ternary Complex Involving a 2:1 Inhibitor/NC Stoichiometry. *J Mol Biol* 2013.
48. Mori M, Schult-Dietrich P, Szafarowicz B, Humbert N, Debaene F, Sanglier-Cianferani S, *et al.* Use of virtual screening for discovering antiretroviral compounds interacting with the HIV-1 nucleocapsid protein. *Virus Res* 2012,**169**:377-387.
49. Nguyen AT, Feasley CL, Jackson KW, Nitz TJ, Salzwedel K, Air GM, *et al.* The prototype HIV-1 maturation inhibitor, bevirimat, binds to the CA-SP1 cleavage site in immature Gag particles. *Retrovirology* 2011,**8**:101.
50. Coric P, Turcaud S, Souquet F, Briant L, Gay B, Royer J, *et al.* Synthesis and biological evaluation of a new derivative of bevirimat that targets the Gag CA-SP1 cleavage site. *Eur J Med Chem* 2013,**62**:453-465.
51. Fodor AA, Aldrich RW. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* 2004,**56**:211-221.
52. Wensing AM, van Maarseveen NM, Nijhuis M. Fifteen years of HIV Protease Inhibitors: raising the barrier to resistance. *Antiviral Res* 2010,**85**:59-74.
53. Blanco JL, Varghese V, Rhee SY, Gatell JM, Shafer RW. HIV-1 integrase inhibitor resistance and its clinical implications. *J Infect Dis* 2011,**203**:1204-1214.
54. Rabi SA, Laird GM, Durand CM, Laskey S, Shan L, Bailey JR, *et al.* Multi-step inhibition explains HIV-1 protease inhibitor pharmacodynamics and resistance. *J Clin Invest* 2013,**123**:3848-3860.
55. Tedbury PR, Ablan SD, Freed EO. Global Rescue of Defects in HIV-1 Envelope Glycoprotein Incorporation: Implications for Matrix Structure. *PLoS Pathog* 2013,**9**:e1003739.
56. Aoki M, Venzon DJ, Koh Y, Aoki-Ogata H, Miyakawa T, Yoshimura K, *et al.* Non-cleavage site gag mutations in amprenavir-resistant human immunodeficiency virus type 1 (HIV-1) predispose HIV-1 to rapid acquisition of amprenavir resistance but delay development of resistance to other protease inhibitors. *J Virol* 2009,**83**:3059-3068.
57. Chang MW, Oliveira G, Yuan J, Okulicz JF, Levy S, Torbett BE. Rapid deep sequencing of patient-derived HIV with ion semiconductor technology. *J Virol Methods* 2013,**189**:232-234.
58. Dussupt V, Sette P, Bello NF, Javid MP, Nagashima K, Bouamr F. Basic residues in the nucleocapsid domain of Gag are critical for late events of HIV-1 budding. *J Virol* 2011,**85**:2304-2315.

59. Taylor JE, Chow JY, Jeffries CM, Kwan AH, Duff AP, Hamilton WA, *et al.* Calmodulin binds a highly extended HIV-1 MA protein that refolds upon its release. *Biophys J* 2012,**103**:541-549.
60. Zuo T, Liu D, Lv W, Wang X, Wang J, Lv M, *et al.* Small-molecule inhibition of human immunodeficiency virus type 1 replication by targeting the interaction between Vif and ElonginC. *J Virol* 2012,**86**:5497-5507.
61. Parry CM, Kolli M, Myers RE, Cane PA, Schiffer C, Pillay D. Three residues in HIV-1 matrix contribute to protease inhibitor susceptibility and replication capacity. *Antimicrob Agents Chemother* 2011,**55**:1106-1113.
62. Yates PJ, Hazen R, St Clair M, Boone L, Tisdale M, Elston RC. In vitro development of resistance to human immunodeficiency virus protease inhibitor GW640385. *Antimicrob Agents Chemother* 2006,**50**:1092-1095.
63. Prado JG, Wrin T, Beauchaine J, Ruiz L, Petropoulos CJ, Frost SD, *et al.* Amprenavir-resistant HIV-1 exhibits lopinavir cross-resistance and reduced replication capacity. *AIDS* 2002,**16**:1009-1017.
64. Mo H, Parkin N, Stewart KD, Lu L, Dekhtyar T, Kempf DJ, *et al.* Identification and structural characterization of I84C and I84A mutations that are associated with high-level resistance to human immunodeficiency virus protease inhibitors and impair viral replication. *Antimicrob Agents Chemother* 2007,**51**:732-735.
65. Nijhuis M, Wensing AM, Bierman WF, de Jong D, Kagan R, Fun A, *et al.* Failure of treatment with first-line lopinavir boosted with ritonavir can be explained by novel resistance pathways with protease mutation 76V. *J Infect Dis* 2009,**200**:698-709.
66. Shibata J, Sugiura W, Ode H, Iwatani Y, Sato H, Tsang H, *et al.* Within-host co-evolution of Gag P453L and protease D30N/N88D demonstrates virological advantage in a highly protease inhibitor-exposed HIV-1 case. *Antiviral Res* 2011,**90**:33-41.
67. Kameoka M, Isarangkura-na-Ayuthaya P, Kameoka Y, Sapsutthipas S, Soonthornsata B, Nakamura S, *et al.* The role of lysine residue at amino acid position 165 of human immunodeficiency virus type 1 CRF01_AE Gag in reducing viral drug susceptibility to protease inhibitors. *Virology* 2010,**405**:129-138.
68. Gatanaga H, Suzuki Y, Tsang H, Yoshimura K, Kavlick MF, Nagashima K, *et al.* Amino acid substitutions in Gag protein at non-cleavage sites are indispensable for the development of a high multitude of HIV-1 resistance against protease inhibitors. *J Biol Chem* 2002,**277**:5952-5961.
69. Dam E, Quercia R, Glass B, Descamps D, Launay O, Duval X, *et al.* Gag mutations strongly contribute to HIV-1 resistance to protease inhibitors in highly drug-experienced patients besides compensating for fitness loss. *PLoS Pathog* 2009,**5**:e1000345.
70. Ho SK, Coman RM, Bunger JC, Rose SL, O'Brien P, Munoz I, *et al.* Drug-associated changes in amino acid residues in Gag p2, p7(NC), and p6(Gag)/p6(Pol) in human immunodeficiency virus type 1 (HIV-1) display a dominant effect on replicative fitness and drug response. *Virology* 2008,**378**:272-281.
71. Brann TW, Dewar RL, Jiang MK, Shah A, Nagashima K, Metcalf JA, *et al.* Functional correlation between a novel amino acid insertion at codon 19 in the protease of human immunodeficiency virus type 1 and polymorphism in the p1/p6 Gag cleavage site in drug resistance and replication fitness. *J Virol* 2006,**80**:6136-6145.
72. Maguire MF, Guinea R, Griffin P, Macmanus S, Elston RC, Wolfram J, *et al.* Changes in human immunodeficiency virus type 1 Gag at positions L449 and P453 are linked to I50V protease mutants in vivo and cause reduction of sensitivity to amprenavir and improved viral fitness in vitro. *J Virol* 2002,**76**:7398-7406.
73. van Maarseveen NM, Andersson D, Lepsik M, Fun A, Schipper PJ, de Jong D, *et al.* Modulation of HIV-1 Gag NC/p1 cleavage efficiency affects protease inhibitor resistance and viral replicative capacity. *Retrovirology* 2012,**9**:29.
74. Kolli M, Stawiski E, Chappey C, Schiffer CA. Human immunodeficiency virus type 1 protease-correlated cleavage site mutations enhance inhibitor resistance. *J Virol* 2009,**83**:11027-11042.
75. Banke S, Lillemark MR, Gerstoft J, Obel N, Jorgensen LB. Positive selection pressure introduces secondary mutations at Gag cleavage sites in human immunodeficiency virus type 1 harboring major protease resistance mutations. *J Virol* 2009,**83**:8916-8924.
76. Bally F, Martinez R, Peters S, Sudre P, Telenti A. Polymorphism of HIV type 1 gag p7/p1 and p1/p6 cleavage sites: clinical significance and implications for resistance to protease inhibitors. *AIDS Res Hum Retroviruses* 2000,**16**:1209-1213.

77. Ghosn J, Delaugerre C, Flandre P, Galimand J, Cohen-Codar I, Raffi F, *et al.* Polymorphism in Gag gene cleavage sites of HIV-1 non-B subtype and virological outcome of a first-line lopinavir/ritonavir single drug regimen. *PLoS One* 2011,**6**:e24798.
78. Lastere S, Dalban C, Collin G, Descamps D, Girard PM, Clavel F, *et al.* Impact of insertions in the HIV-1 p6 PTAPP region on the virological response to amprenavir. *Antivir Ther* 2004,**9**:221-227.
79. Kaufmann GR, Suzuki K, Cunningham P, Mukaide M, Kondo M, Imai M, *et al.* Impact of HIV type 1 protease, reverse transcriptase, cleavage site, and p6 mutations on the virological response to quadruple therapy with saquinavir, ritonavir, and two nucleoside analogs. *AIDS Res Hum Retroviruses* 2001,**17**:487-497.
80. Lambert-Niclot S, Flandre P, Malet I, Canestri A, Soulie C, Tubiana R, *et al.* Impact of gag mutations on selection of darunavir resistance mutations in HIV-1 protease. *J Antimicrob Chemother* 2008,**62**:905-908.
81. Cote HC, Brumme ZL, Harrigan PR. Human immunodeficiency virus type 1 protease cleavage site mutations associated with protease inhibitor cross-resistance selected by indinavir, ritonavir, and/or saquinavir. *J Virol* 2001,**75**:589-594.
82. Larrouy L, Chazallon C, Landman R, Capitant C, Peytavin G, Collin G, *et al.* Gag mutations can impact virological response to dual-boosted protease inhibitor combinations in antiretroviral-naive HIV-infected patients. *Antimicrob Agents Chemother* 2010,**54**:2910-2919.
83. Larrouy L, Charpentier C, Landman R, Capitant C, Chazallon C, Yeni P, *et al.* Dynamics of gag-pol minority viral populations in naive HIV-1-infected patients failing protease inhibitor regimen. *AIDS* 2011,**25**:2143-2148.
84. Zhang YM, Imamichi H, Imamichi T, Lane HC, Falloon J, Vasudevachari MB, *et al.* Drug resistance during indinavir therapy is caused by mutations in the protease gene and in its Gag substrate cleavage sites. *J Virol* 1997,**71**:6662-6670.
85. Roquebert B, Malet I, Wirden M, Tubiana R, Valantin MA, Simon A, *et al.* Role of HIV-1 minority populations on resistance mutational pattern evolution and susceptibility to protease inhibitors. *AIDS* 2006,**20**:287-289.
86. Verheyen J, Litau E, Sing T, Daumer M, Balduin M, Oette M, *et al.* Compensatory mutations at the HIV cleavage sites p7/p1 and p1/p6-gag in therapy-naive and therapy-experienced patients. *Antivir Ther* 2006,**11**:879-887.
87. Malet I, Roquebert B, Dalban C, Wirden M, Amellal B, Agher R, *et al.* Association of Gag cleavage sites to protease mutations and to virological response in HIV-1 treated patients. *J Infect* 2007,**54**:367-374.
88. Rossi AH, Rocco CA, Mangano A, Sen L, Aulicino PC. Sequence variability in p6 gag protein and gag/pol coevolution in human immunodeficiency type 1 subtype F genomes. *AIDS Res Hum Retroviruses* 2013,**29**:1056-1060.
89. Feher A, Weber IT, Bagossi P, Boross P, Mahalingam B, Louis JM, *et al.* Effect of sequence polymorphism and drug resistance on two HIV-1 Gag processing sites. *Eur J Biochem* 2002,**269**:4114-4120.
90. Peters S, Munoz M, Yerly S, Sanchez-Merino V, Lopez-Galindez C, Perrin L, *et al.* Resistance to nucleoside analog reverse transcriptase inhibitors mediated by human immunodeficiency virus type 1 p6 protein. *J Virol* 2001,**75**:9644-9653.
91. Flys T, Marlowe N, Hackett J, Parkin N, Schumaker M, Holzmayer V, *et al.* Analysis of PTAP duplications in the gag p6 region of subtype C HIV type 1. *AIDS Res Hum Retroviruses* 2005,**21**:739-741.
92. Gallego O, de Mendoza C, Corral A, Soriano V. Changes in the human immunodeficiency virus p7-p1-p6 gag gene in drug-naive and pretreated patients. *J Clin Microbiol* 2003,**41**:1245-1247.
93. Martins AN, Arruda MB, Pires AF, Tanuri A, Brindeiro RM. Accumulation of P(T/S)AP late domain duplications in HIV type 1 subtypes B, C, and F derived from individuals failing ARV therapy and ARV drug-naive patients. *AIDS Res Hum Retroviruses* 2011,**27**:687-692.
94. Brumme ZL, Chan KJ, Dong WW, Wynhoven B, Mo T, Hogg RS, *et al.* Prevalence and clinical implications of insertions in the HIV-1 p6Gag N-terminal region in drug-naive individuals initiating antiretroviral therapy. *Antivir Ther* 2003,**8**:91-96.
95. Pettit SC, Moody MD, Wehbie RS, Kaplan AH, Nantermet PV, Klein CA, *et al.* The p2 domain of human immunodeficiency virus type 1 Gag regulates sequential proteolytic processing and is required to produce fully infectious virions. *J Virol* 1994,**68**:8017-8027.

96. Verheyen J, Knops E, Kupfer B, Hamouda O, Somogyi S, Schuldenzucker U, *et al.* Prevalence of C-terminal gag cleavage site mutations in HIV from therapy-naive patients. *J Infect* 2009,**58**:61-67.
97. Adamson CS, Sakalian M, Salzwedel K, Freed EO. Polymorphisms in Gag spacer peptide 1 confer varying levels of resistance to the HIV- 1 maturation inhibitor bevirimat. *Retrovirology* 2010,**7**:36.
98. Van Baelen K, Salzwedel K, Rondelez E, Van Eygen V, De Vos S, Verheyen A, *et al.* Susceptibility of human immunodeficiency virus type 1 to the maturation inhibitor bevirimat is modulated by baseline polymorphisms in Gag spacer peptide 1. *Antimicrob Agents Chemother* 2009,**53**:2185-2188.
99. Seclen E, Gonzalez Mdel M, Corral A, de Mendoza C, Soriano V, Poveda E. High prevalence of natural polymorphisms in Gag (CA-SP1) associated with reduced response to Bevirimat, an HIV-1 maturation inhibitor. *AIDS* 2010,**24**:467-469.
100. Huang L, Li Y, Chen C. Flexible catalytic site conformations implicated in modulation of HIV-1 protease autoprocessing reactions. *Retrovirology* 2011,**8**:79.
101. Lee SK, Harris J, Swanstrom R. A strongly transdominant mutation in the human immunodeficiency virus type 1 gag gene defines an Achilles heel in the virus life cycle. *J Virol* 2009,**83**:8536-8543.
102. Thomas JA, Shulenin S, Coren LV, Bosche WJ, Gagliardi TD, Gorelick RJ, *et al.* Characterization of human immunodeficiency virus type 1 (HIV-1) containing mutations in the nucleocapsid protein at a putative HIV-1 protease cleavage site. *Virology* 2006,**354**:261-270.
103. Zhou J, Chen CH, Aiken C. Human immunodeficiency virus type 1 resistance to the small molecule maturation inhibitor 3-O-(3',3'-dimethylsuccinyl)-betulinic acid is conferred by a variety of single amino acid substitutions at the CA-SP1 cleavage site in Gag. *J Virol* 2006,**80**:12095-12101.
104. Tamiya S, Mardy S, Kavlick MF, Yoshimura K, Mistuya H. Amino acid insertions near Gag cleavage sites restore the otherwise compromised replication of human immunodeficiency virus type 1 variants resistant to protease inhibitors. *J Virol* 2004,**78**:12030-12040.
105. de Oliveira T, Engelbrecht S, Janse van Rensburg E, Gordon M, Bishop K, zur Megede J, *et al.* Variability at human immunodeficiency virus type 1 subtype C protease cleavage sites: an indication of viral fitness? *J Virol* 2003,**77**:9422-9430.
106. Goodenow MM, Bloom G, Rose SL, Pomeroy SM, O'Brien PO, Perez EE, *et al.* Naturally occurring amino acid polymorphisms in human immunodeficiency virus type 1 (HIV-1) Gag p7(NC) and the C-cleavage site impact Gag-Pol processing by HIV-1 protease. *Virology* 2002,**292**:137-149.
107. Mammano F, Petit C, Clavel F. Resistance-associated loss of viral fitness in human immunodeficiency virus type 1: phenotypic analysis of protease and gag coevolution in protease inhibitor-treated patients. *J Virol* 1998,**72**:7632-7637.
108. McKinnon JE, Delgado R, Pulido F, Shao W, Arribas JR, Mellors JW. Single genome sequencing of HIV-1 gag and protease resistance mutations at virologic failure during the OK04 trial of simplified versus standard maintenance therapy. *Antivir Ther* 2011,**16**:725-732.
109. Resch W, Parkin N, Watkins T, Harris J, Swanstrom R. Evolution of human immunodeficiency virus type 1 protease genotypes and phenotypes in vivo under selective pressure of the protease inhibitor ritonavir. *J Virol* 2005,**79**:10638-10649.
110. Parry CM, Kohli A, Boinett CJ, Towers GJ, McCormick AL, Pillay D. Gag determinants of fitness and drug susceptibility in protease inhibitor-resistant human immunodeficiency virus type 1. *J Virol* 2009,**83**:9094-9101.
111. Ho SK, Perez EE, Rose SL, Coman RM, Lowe AC, Hou W, *et al.* Genetic determinants in HIV-1 Gag and Env V3 are related to viral response to combination antiretroviral therapy with a protease inhibitor. *AIDS* 2009,**23**:1631-1640.
112. Matsuoka-Aizawa S, Gatanaga H, Sato H, Koike K, Kimura S, Oka S. Cooperative contribution of gag substitutions to nelfinavir-dependent enhancement of precursor cleavage and replication of human immunodeficiency virus type-1. *Antiviral Res* 2006,**70**:51-59.
113. Samal AB, Ghanam RH, Fernandez TF, Monroe EB, Saad JS. NMR, biophysical, and biochemical studies reveal the minimal Calmodulin binding domain of the HIV-1 matrix protein. *J Biol Chem* 2011,**286**:33533-33543.
114. Chow JY, Jeffries CM, Kwan AH, Guss JM, Trewhella J. Calmodulin disrupts the structure of the HIV-1 MA protein. *J Mol Biol* 2010,**400**:702-714.

115. Ghanam RH, Fernandez TF, Fledderman EL, Saad JS. Binding of calmodulin to the HIV-1 matrix protein triggers myristate exposure. *J Biol Chem* 2010,**285**:41911-41920.
116. Peytavi R, Hong SS, Gay B, d'Angeac AD, Selig L, Benichou S, *et al.* HEED, the product of the human homolog of the murine eed gene, binds to the matrix protein of HIV-1. *J Biol Chem* 1999,**274**:1635-1645.
117. Rakotobe D, Violot S, Hong SS, Gouet P, Boulanger P. Mapping of immunogenic and protein-interacting regions at the surface of the seven-bladed beta-propeller domain of the HIV-1 cellular interactor EED. *Virol J* 2008,**5**:32.
118. Giagulli C, Magiera AK, Bugatti A, Caccuri F, Marsico S, Rusnati M, *et al.* HIV-1 matrix protein p17 binds to the IL-8 receptor CXCR1 and shows IL-8-like chemokine activity on monocytes through Rho/ROCK activation. *Blood* 2012,**119**:2274-2283.
119. Caccuri F, Giagulli C, Bugatti A, Benetti A, Alessandri G, Ribatti D, *et al.* HIV-1 matrix protein p17 promotes angiogenesis via chemokine receptors CXCR1 and CXCR2. *Proc Natl Acad Sci U S A* 2012,**109**:14580-14585.
120. Bugatti A, Giagulli C, Urbinati C, Caccuri F, Chiodelli P, Oreste P, *et al.* Molecular interaction studies of HIV-1 matrix protein p17 and heparin: identification of the heparin-binding motif of p17 as a target for the development of multitarget antagonists. *J Biol Chem* 2013,**288**:1150-1161.
121. Cimarelli A, Luban J. Translation elongation factor 1-alpha interacts specifically with the human immunodeficiency virus type 1 Gag polyprotein. *J Virol* 1999,**73**:5388-5401.
122. Bristow R, Byrne J, Squirell J, Trencher H, Carter T, Rodgers B, *et al.* Human cyclophilin has a significantly higher affinity for HIV-1 recombinant p55 than p24. *J Acquir Immune Defic Syndr Hum Retrovirol* 1999,**20**:334-336.
123. Dupont S, Sharova N, DeHoratius C, Virbasius CM, Zhu X, Bukrinskaya AG, *et al.* A novel nuclear export activity in HIV-1 matrix protein required for viral replication. *Nature* 1999,**402**:681-685.
124. Haffar OK, Popov S, Dubrovsky L, Agostini I, Tang H, Pushkarsky T, *et al.* Two nuclear localization signals in the HIV-1 matrix protein regulate nuclear import of the HIV-1 pre-integration complex. *J Mol Biol* 2000,**299**:359-368.
125. Batonick M, Favre M, Boge M, Spearman P, Honing S, Thali M. Interaction of HIV-1 Gag with the clathrin-associated adaptor AP-2. *Virology* 2005,**342**:190-200.
126. Lopez-Verges S, Camus G, Blot G, Beauvoir R, Benarous R, Berlioz-Torrent C. Tail-interacting protein TIP47 is a connector between Gag and Env and is required for Env incorporation into HIV-1 virions. *Proc Natl Acad Sci U S A* 2006,**103**:14947-14952.
127. Bauby H, Lopez-Verges S, Hoeffel G, Delcroix-Genete D, Janvier K, Mammano F, *et al.* TIP47 is required for the production of infectious HIV-1 particles from primary macrophages. *Traffic* 2010,**11**:455-467.
128. Lama J, Trono D. Human immunodeficiency virus type 1 matrix protein interacts with cellular protein HO3. *J Virol* 1998,**72**:1671-1676.
129. Kaushik R, Ratner L. Role of human immunodeficiency virus type 1 matrix phosphorylation in an early postentry step of virus replication. *J Virol* 2004,**78**:2319-2326.
130. Jacque JM, Mann A, Enslen H, Sharova N, Brichacek B, Davis RJ, *et al.* Modulation of HIV-1 infectivity by MAPK, a virion-associated kinase. *EMBO J* 1998,**17**:2607-2618.
131. Gupta P, Singhal PK, Rajendrakumar P, Padwad Y, Tendulkar AV, Kalyanaraman VS, *et al.* Mechanism of host cell MAPK/ERK-2 incorporation into lentivirus particles: characterization of the interaction between MAPK/ERK-2 and proline-rich-domain containing capsid region of structural protein Gag. *J Mol Biol* 2011,**410**:681-697.
132. Vlach J, Saad JS. Trio engagement via plasma membrane phospholipids and the myristoyl moiety governs HIV-1 matrix binding to bilayers. *Proc Natl Acad Sci U S A* 2013,**110**:3525-3530.
133. Steckbeck JD, Kuhlmann AS, Montelaro RC. C-terminal tail of human immunodeficiency virus gp41: functionally rich and structurally enigmatic. *J Gen Virol* 2013,**94**:1-19.
134. Lin CW, Engelman A. The barrier-to-autointegration factor is a component of functional human immunodeficiency virus type 1 preintegration complexes. *J Virol* 2003,**77**:5030-5036.
135. Burnette B, Yu G, Felsted RL. Phosphorylation of HIV-1 gag proteins by protein kinase C. *J Biol Chem* 1993,**268**:8698-8703.
136. Yu G, Shen FS, Sturch S, Aquino A, Glazer RI, Felsted RL. Regulation of HIV-1 gag protein subcellular targeting by protein kinase C. *J Biol Chem* 1995,**270**:4792-4796.
137. !!! INVALID CITATION !!!

138. Buratti E, Tisminetzky SG, D'Agaro P, Baralle FE. A neutralizing monoclonal antibody previously mapped exclusively on human immunodeficiency virus type 1 gp41 recognizes an epitope in p17 sharing the core sequence IEEE. *J Virol* 1997,**71**:2457-2462.
139. Saad JS, Kim A, Ghanam RH, Dalton AK, Vogt VM, Wu Z, *et al.* Mutations that mimic phosphorylation of the HIV-1 matrix protein do not perturb the myristyl switch. *Protein Sci* 2007,**16**:1793-1797.
140. Saad JS, Miller J, Tai J, Kim A, Ghanam RH, Summers MF. Structural basis for targeting HIV-1 Gag proteins to the plasma membrane for virus assembly. *Proc Natl Acad Sci U S A* 2006,**103**:11364-11369.
141. Saad JS, Ablan SD, Ghanam RH, Kim A, Andrews K, Nagashima K, *et al.* Structure of the myristylated human immunodeficiency virus type 2 matrix protein and the role of phosphatidylinositol-(4,5)-bisphosphate in membrane targeting. *J Mol Biol* 2008,**382**:434-447.
142. Gamble TR, Vajdos FF, Yoo S, Worthylake DK, Houseweart M, Sundquist WI, *et al.* Crystal structure of human cyclophilin A bound to the amino-terminal domain of HIV-1 capsid. *Cell* 1996,**87**:1285-1294.
143. Qi M, Yang R, Aiken C. Cyclophilin A-dependent restriction of human immunodeficiency virus type 1 capsid mutants for infection of nondividing cells. *J Virol* 2008,**82**:12001-12008.
144. Ylinen LM, Schaller T, Price A, Fletcher AJ, Noursadeghi M, James LC, *et al.* Cyclophilin A levels dictate infection efficiency of human immunodeficiency virus type 1 capsid escape mutants A92E and G94D. *J Virol* 2009,**83**:2044-2047.
145. Gatanaga H, Das D, Suzuki Y, Yeh DD, Hussain KA, Ghosh AK, *et al.* Altered HIV-1 Gag protein interactions with cyclophilin A (CypA) on the acquisition of H219Q and H219P substitutions in the CypA binding loop. *J Biol Chem* 2006,**281**:1241-1250.
146. Colgan J, Yuan HE, Franke EK, Luban J. Binding of the human immunodeficiency virus type 1 Gag polyprotein to cyclophilin A is mediated by the central region of capsid and requires Gag dimerization. *J Virol* 1996,**70**:4299-4310.
147. Ambrose Z, Lee K, Ndjomou J, Xu H, Oztop I, Matous J, *et al.* Human immunodeficiency virus type 1 capsid mutation N74D alters cyclophilin A dependence and impairs macrophage infection. *J Virol* 2012,**86**:4708-4714.
148. Mascarenhas AP, Musier-Forsyth K. The capsid protein of human immunodeficiency virus: interactions of HIV-1 capsid with host protein factors. *FEBS J* 2009,**276**:6118-6127.
149. Yoo S, Myszka DG, Yeh C, McMurray M, Hill CP, Sundquist WI. Molecular recognition in the HIV-1 capsid/cyclophilin A complex. *J Mol Biol* 1997,**269**:780-795.
150. Song C, Aiken C. Analysis of human cell heterokaryons demonstrates that target cell restriction of cyclosporine-resistant human immunodeficiency virus type 1 mutants is genetically dominant. *J Virol* 2007,**81**:11946-11956.
151. Hatziioannou T, Perez-Caballero D, Cowan S, Bieniasz PD. Cyclophilin interactions with incoming human immunodeficiency virus type 1 capsids with opposing effects on infectivity in human cells. *J Virol* 2005,**79**:176-183.
152. Price AJ, Fletcher AJ, Schaller T, Elliott T, Lee K, KewalRamani VN, *et al.* CPSF6 defines a conserved capsid interface that modulates HIV-1 replication. *PLoS Pathog* 2012,**8**:e1002896.
153. Lee K, Ambrose Z, Martin TD, Oztop I, Mulky A, Julias JG, *et al.* Flexible use of nuclear import pathways by HIV-1. *Cell Host Microbe* 2010,**7**:221-233.
154. Sayah DM, Sokolskaja E, Berthoux L, Luban J. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature* 2004,**430**:569-573.
155. Stremlau M, Perron M, Lee M, Li Y, Song B, Javanbakht H, *et al.* Specific recognition and accelerated uncoating of retroviral capsids by the TRIM5alpha restriction factor. *Proc Natl Acad Sci U S A* 2006,**103**:5514-5519.
156. Yang R, Shi J, Byeon IJ, Ahn J, Sheehan JH, Meiler J, *et al.* Second-site suppressors of HIV-1 capsid mutations: restoration of intracellular activities without correction of intrinsic capsid stability defects. *Retrovirology* 2012,**9**:30.
157. Maillard PV, Zoete V, Michielin O, Trono D. Homology-based identification of capsid determinants that protect HIV1 from human TRIM5alpha restriction. *J Biol Chem* 2011,**286**:8128-8140.
158. Diaz-Griffero F, Qin XR, Hayashi F, Kigawa T, Finzi A, Sarnak Z, *et al.* A B-box 2 surface patch important for TRIM5alpha self-association, capsid binding avidity, and retrovirus restriction. *J Virol* 2009,**83**:10737-10751.
159. Stremlau M, Owens CM, Perron MJ, Kiessling M, Autissier P, Sodroski J. The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys. *Nature* 2004,**427**:848-853.

160. Li X, Song B, Xiang SH, Sodroski J. Functional interplay between the B-box 2 and the B30.2(SPRY) domains of TRIM5alpha. *Virology* 2007,**366**:234-244.
161. Li Y, Li X, Stremlau M, Lee M, Sodroski J. Removal of arginine 332 allows human TRIM5alpha to bind human immunodeficiency virus capsids and to restrict infection. *J Virol* 2006,**80**:6738-6744.
162. De Iaco A, Luban J. Inhibition of HIV-1 infection by TNPO3 depletion is determined by capsid and detectable after viral cDNA enters the nucleus. *Retrovirology* 2011,**8**:98.
163. Zhou L, Sokolskaja E, Jolly C, James W, Cowley SA, Fassati A. Transportin 3 promotes a nuclear maturation step required for efficient HIV-1 integration. *PLoS Pathog* 2011,**7**:e1002194.
164. Valle-Casuso JC, Di Nunzio F, Yang Y, Reszka N, Lienlaf M, Arhel N, *et al.* TNPO3 is required for HIV-1 replication after nuclear import but prior to integration and binds the HIV-1 core. *J Virol* 2012,**86**:5931-5936.
165. Di Nunzio F, Danckaert A, Fricke T, Perez P, Fernandez J, Perret E, *et al.* Human nucleoporins promote HIV-1 docking at the nuclear pore, nuclear import and integration. *PLoS One* 2012,**7**:e46037.
166. Di Nunzio F, Fricke T, Miccio A, Valle-Casuso JC, Perez P, Souque P, *et al.* Nup153 and Nup98 bind the HIV-1 core and contribute to the early steps of HIV-1 replication. *Virology* 2013,**440**:8-18.
167. Matreyek KA, Yucel SS, Li X, Engelman A. Nucleoporin NUP153 Phenylalanine-Glycine Motifs Engage a Common Binding Pocket within the HIV-1 Capsid Protein to Mediate Lentiviral Infectivity. *PLoS Pathog* 2013,**9**:e1003693.
168. Matreyek KA, Engelman A. The requirement for nucleoporin NUP153 during human immunodeficiency virus type 1 infection is determined by the viral capsid. *J Virol* 2011,**85**:7818-7827.
169. Koh Y, Wu X, Ferris AL, Matreyek KA, Smith SJ, Lee K, *et al.* Differential effects of human immunodeficiency virus type 1 capsid and cellular factors nucleoporin 153 and LEDGF/p75 on the efficiency and specificity of viral DNA integration. *J Virol* 2013,**87**:648-658.
170. Bichel K, Price AJ, Schaller T, Towers GJ, Freund SM, James LC. HIV-1 capsid undergoes coupled binding and isomerization by the nuclear pore protein NUP358. *Retrovirology* 2013,**10**:81.
171. Ocwieja KE, Brady TL, Ronen K, Huegel A, Roth SL, Schaller T, *et al.* HIV integration targeting: a pathway involving Transportin-3 and the nuclear pore protein RanBP2. *PLoS Pathog* 2011,**7**:e1001313.
172. Schaller T, Ocwieja KE, Rasaiyaah J, Price AJ, Brady TL, Roth SL, *et al.* HIV-1 capsid-cyclophilin interactions determine nuclear import pathway, integration targeting and replication efficiency. *PLoS Pathog* 2011,**7**:e1002439.
173. Misumi S, Inoue M, Dochi T, Kishimoto N, Hasegawa N, Takamune N, *et al.* Uncoating of human immunodeficiency virus type 1 requires prolyl isomerase Pin1. *J Biol Chem* 2010,**285**:25185-25195.
174. Kitagawa Y, Kameoka M, Shoji-Kawata S, Iwabu Y, Mizuta H, Tokunaga K, *et al.* Inhibitory function of adapter-related protein complex 2 alpha 1 subunit in the process of nuclear translocation of human immunodeficiency virus type 1 genome. *Virology* 2008,**373**:171-180.
175. Javanbakht H, Halwani R, Cen S, Saadatmand J, Musier-Forsyth K, Gottlinger H, *et al.* The interaction between HIV-1 Gag and human lysyl-tRNA synthetase during viral assembly. *J Biol Chem* 2003,**278**:27644-27651.
176. Abudu A, Wang X, Dang Y, Zhou T, Xiang SH, Zheng YH. Identification of molecular determinants from Moloney leukemia virus 10 homolog (MOV10) protein for virion packaging and anti-HIV-1 activity. *J Biol Chem* 2012,**287**:1220-1228.
177. Wang X, Han Y, Dang Y, Fu W, Zhou T, Ptak RG, *et al.* Moloney leukemia virus 10 (MOV10) protein inhibits retrovirus replication. *J Biol Chem* 2010,**285**:14346-14355.
178. Dussupt V, Javid MP, Abou-Jaoude G, Jadwin JA, de La Cruz J, Nagashima K, *et al.* The nucleocapsid region of HIV-1 Gag cooperates with the PTAP and LYPXnL late domains to recruit the cellular machinery necessary for viral budding. *PLoS Pathog* 2009,**5**:e1000339.
179. Sette P, Jadwin JA, Dussupt V, Bello NF, Bouamr F. The ESCRT-associated protein Alix recruits the ubiquitin ligase Nedd4-1 to facilitate HIV-1 release through the LYPXnL L domain motif. *J Virol* 2010,**84**:8181-8192.
180. Zhai Q, Landesman MB, Robinson H, Sundquist WI, Hill CP. Structure of the Bro1 domain protein BROX and functional analyses of the ALIX Bro1 domain in HIV-1 budding. *PLoS One* 2011,**6**:e27466.

181. Sette P, Dussupt V, Bouamr F. Identification of the HIV-1 NC binding interface in Alix Bro1 reveals a role for RNA. *J Virol* 2012,**86**:11608-11615.
182. Luo K, Liu B, Xiao Z, Yu Y, Yu X, Gorelick R, *et al.* Amino-terminal region of the human immunodeficiency virus type 1 nucleocapsid is required for human APOBEC3G packaging. *J Virol* 2004,**78**:11841-11852.
183. Burnett A, Spearman P. APOBEC3G multimers are recruited to the plasma membrane for packaging into human immunodeficiency virus type 1 virus-like particles in an RNA-dependent process requiring the NC basic linker. *J Virol* 2007,**81**:5000-5013.
184. Cen S, Guo F, Niu M, Saadatmand J, Deflassieux J, Kleiman L. The interaction between HIV-1 Gag and APOBEC3G. *J Biol Chem* 2004,**279**:33177-33184.
185. Huthoff H, Malim MH. Identification of amino acid residues in APOBEC3G required for regulation by human immunodeficiency virus type 1 Vif and Virion encapsidation. *J Virol* 2007,**81**:3807-3815.
186. Zhou Y, Rong L, Lu J, Pan Q, Liang C. Insulin-like growth factor II mRNA binding protein 1 associates with Gag protein of human immunodeficiency virus type 1, and its overexpression affects virus assembly. *J Virol* 2008,**82**:5683-5692.
187. Chatel-Chaix L, Boulay K, Mouland AJ, Desgroseillers L. The host protein Staufen1 interacts with the Pr55Gag zinc fingers and regulates HIV-1 assembly via its N-terminus. *Retrovirology* 2008,**5**:41.
188. Lingappa JR, Dooher JE, Newman MA, Kiser PK, Klein KC. Basic residues in the nucleocapsid domain of Gag are required for interaction of HIV-1 gag with ABCE1 (HP68), a cellular protein important for HIV-1 capsid assembly. *J Biol Chem* 2006,**281**:3773-3784.
189. Takahashi H, Matsuda M, Kojima A, Sata T, Andoh T, Kurata T, *et al.* Human immunodeficiency virus type 1 reverse transcriptase: enhancement of activity by interaction with cellular topoisomerase I. *Proc Natl Acad Sci U S A* 1995,**92**:5694-5698.
190. Fisher RD, Chung HY, Zhai Q, Robinson H, Sundquist WI, Hill CP. Structural and biochemical studies of ALIX/AIP1 and its role in retrovirus budding. *Cell* 2007,**128**:841-852.
191. Lazert C, Chazal N, Briant L, Gerlier D, Cortay JC. Refined study of the interaction between HIV-1 p6 late domain and ALIX. *Retrovirology* 2008,**5**:39.
192. Patil A, Bhattacharya J. Natural deletion of L35Y36 in p6 gag eliminate LYPXnL/ALIX auxiliary virus release pathway in HIV-1 subtype C. *Virus Res* 2012,**170**:154-158.
193. Strack B, Calistri A, Craig S, Popova E, Gottlinger HG. AIP1/ALIX is a binding partner for HIV-1 p6 and EIAV p9 functioning in virus budding. *Cell* 2003,**114**:689-699.
194. Gurer C, Berthoux L, Luban J. Covalent modification of human immunodeficiency virus type 1 p6 by SUMO-1. *J Virol* 2005,**79**:910-917.
195. Jaber T, Bohl CR, Lewis GL, Wood C, West JT, Jr., Weldon RA, Jr. Human Ubc9 contributes to production of fully infectious human immunodeficiency virus type 1 virions. *J Virol* 2009,**83**:10448-10459.
196. Hemonnot B, Cartier C, Gay B, Rebuffat S, Bardy M, Devaux C, *et al.* The host cell MAP kinase ERK-2 regulates viral assembly and release by phosphorylating the p6gag protein of HIV-1. *J Biol Chem* 2004,**279**:32426-32434.
197. Solbak SM, Reksten TR, Roder R, Wray V, Horvli O, Raae AJ, *et al.* HIV-1 p6-Another viral interaction partner to the host cellular protein cyclophilin A. *Biochim Biophys Acta* 2012,**1824**:667-678.
198. Ott DE, Coren LV, Copeland TD, Kane BP, Johnson DG, Sowder RC, 2nd, *et al.* Ubiquitin is covalently attached to the p6Gag proteins of human immunodeficiency virus type 1 and simian immunodeficiency virus and to the p12Gag protein of Moloney murine leukemia virus. *J Virol* 1998,**72**:2962-2968.
199. Johnson VA, Calvez V, Günthard HF, Paredes R, Pillay D, Shafer RW, *et al.* Update of the Drug Resistance Mutations in HIV-1: March 2013. *Top Antivir Med*,**21**:4-12.
200. Shafer RW. Rationale and Uses of a Public HIV Drug-Resistance Database. *Journal of Infectious Diseases* 2006,**194**:S51-S58.
201. Vercauteren J, Beheydt G, Prosperi M, Libin P, Imbrechts S, Camacho R, *et al.* Clinical evaluation of Rega 8: an updated genotypic interpretation system that significantly predicts HIV-therapy response. *PLoS One* 2013,**8**:e61436.
202. SIDA ANdRsl. ANRS genotypic resistance guidelines (version 22). 2012.
203. Larrouy L, Vivot A, Charpentier C, Benard A, Visseaux B, Damond F, *et al.* Impact of gag genetic determinants on virological outcome to boosted lopinavir-containing regimen in HIV-2-infected patients. *AIDS* 2013,**27**:69-80.

204. Myint L, Matsuda M, Matsuda Z, Yokomaku Y, Chiba T, Okano A, *et al.* Gag non-cleavage site mutations contribute to full recovery of viral fitness in protease inhibitor-resistant human immunodeficiency virus type 1. *Antimicrob Agents Chemother* 2004,**48**:444-452.
205. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* 2009,**19**:1639-1645.
206. Rose PP, Korber BT. Detecting hypermutations in viral sequences with an emphasis on G --> A hypermutation. *Bioinformatics* 2000,**16**:400-401.
207. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004,**32**:1792-1797.
208. Garcıa S, Herrera F. An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons. *Journal of Machine Learning Research* 2008,**9**:66.
209. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 2011,**27**:1164-1165.
210. DeLano WL. The PyMOL molecular graphics system. 2002.
211. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The binormal assumption on precision-recall curves. In: *Pattern Recognition (ICPR), 2010 20th International Conference on*: IEEE; 2010. pp. 4263-4266.
212. Klose DP, Wallace BA, Janes RW. 2Struc: the secondary structure server. *Bioinformatics* 2010,**26**:2624-2625.
213. Stamatakis A, Aberer AJ, Goll C, Smith SA, Berger SA, Izquierdo-Carrasco F. RAxML-Light: a tool for computing terabyte phylogenies. *Bioinformatics* 2012,**28**:2064-2066.

Chapter 7

Learning ancestral polytrees

"Genius is one percent inspiration and ninety-nine percent perspiration."

—Thomas Edison

This chapter is an adapted reprint of my article:

Guangdi Li, Anne-Mieke Vandamme, Jan Ramon, Learning Ancestral Polytrees. The 1st workshop of Learning Tractable Probabilistic Model at the 31st International Conference on Machine Learning (ICML), 2014

I proposed the idea, performed the analysis, designed the software and drafted the manuscript. The improvement of the paper was supported with substantial help from Prof. Jan Ramon and Prof. Anne-Mieke Vandamme.

7.1 Summary

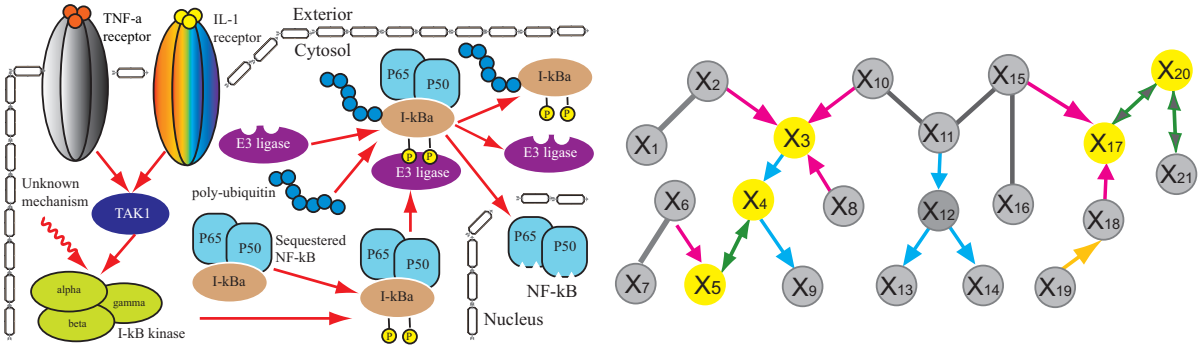
Causal polytrees are singly connected causal models and they are frequently applied in practice. However, in various applications, many variables remain unobserved and causal polytrees cannot be applied without explicitly including unobserved variables. Our study thus proposes the ancestral polytree model, a novel combination of ancestral graphs and singly connected graphs. Ancestral graphs can model causal and non-causal dependencies, while singly connected models allow for efficient learning and inference. We discuss the basic properties of ancestral polytrees and propose an efficient structure learning algorithm. Experiments on synthetic datasets and biological datasets show that our algorithm is efficient and the applications of ancestral polytrees are promising.

7.2 Introduction

Causal graphical models have been proposed to explicitly convey causal relations between causes and their effects in reasoning tasks [1]. As a special class, polytrees are singly connected graphical models where each pair of variables is connected through at most one path [2]. Since Rebane and Pearl introduced polytree-like Bayesian networks [3], called dependency polytrees, further research has shown that, (1) belief propagation can be performed computationally efficiently in polytrees [4]. (2) Learning the maximum likelihood dependency polytrees was proven to be NP-hard [5]. (3) Polynomial algorithms were proposed to learn causal polytrees via conditional independence (CI) tests [2, 6]. (4) Based on independency properties of isomorphic polytrees [7], a sound and complete criterion was proposed to read independence relations from minimal directed independence maps [8].

Insofar, many variables remain unobserved in many applications, which have driven us to design robust causal models bearing unobserved (synonymous with latent and hidden) variables. However, learning large Bayesian networks is slow [6] and causal polytrees cannot express causal flows without explicitly including unobserved variables, causing the increased complexity of causal reasoning, structure learning and inference. The drawbacks above motivate us to introduce ancestral polytrees (APs) with a fast structure learning algorithm and we show this new model can be used to learn many biological systems.

Polytree models have been applied in real world applications. For example, dependency polytrees were efficiently implemented to enhance caching strategies in distributed databases [10]. Based on dependency polytrees, an inference framework was successfully designed



(a) An NF-kB protein signaling pathway

(b) A polytree network with 21 vertices

Figure 1. (a)An NF-kB protein signaling pathway (adapted from [9]). IL-1 and TNF-α are transmitted to IL-1 receptors and TNF-α receptors due to extracellular signals. These receptors then activate TAK1, which immediately activates the I-kBα kinase. The I-kBα kinase phosphorylates two serine residues in the I-kBα, allowing for the further binding of the E3 ligase to trigger the degradation of I-kBα and NF-kB. Thereafter, NF-kB is transported into nucleus and activates gene transcription. Twisted red arrow indicates that there is still an unknown mechanism which cooperates with TAK1 to activate the I-kBα kinase. (b) A polytree network with 21 vertices.

to

optimize hardware components according to the performance and price of both traditional and nanotechnology architectures [11]. Moreover, protein signaling pathways might be modeled by causal polytrees. For instance, Figure 1 illustrates an NF-kB protein signaling pathway, which activates mammalian immune system cells to produce antibodies against inflammation [9]. In this example, causal flows are indicated by red arrows and the activation or inhibition of involved proteins represent cause or effect events.

Protein signaling pathways can be aptly modeled in cases where protein signalling data have been collected, for instance, using multiparameter flow cytometry [12]. So far, many proteins in protein signaling pathways remain unobserved, which have driven us to design robust causal models bearing latent variables. However, causal polytrees cannot express causal flows without explicitly invoking latent variables, causing the increased complexity of causal reasonings. Fortunately, ancestral graphs (AGs) have been proposed to model latent variables without invoking any additional variables [13]. This has motivated us to introduce ancestral polytrees (APs) as the extensions of causal polytrees.

In ancestral graphs, every missing edge indicates an independence relation [14]. Besides

the lack of any directed cycles in both DAGs and AGs, AGs contain directed edges, bidirected edges and undirected edges, in contrast to DAGs, which only permit directed edges. AGs are refined surrogates for DAGs even in the presence of unobserved variables and selection effects [15]. Since causal polytrees are the offsprings of DAGs and singly connected networks, APs are the progenies of AGs and singly connected networks. This merit allows for APs to inherit the merits of both AGs and singly connected networks. On the one hand, APs can express causal diagrams without invoking additional variables in the presence of unobserved variables. On the other hand, due to their simplified structures, APs might guarantee a fast structure learning and inference compared to DAGs and general AGs. However, this inheritance in turn demands a strong assumption — that is, the underlying reasoning diagrams of APs must be singly connected.

This paper begins with basic definitions. We then characterize the properties of APs regarding Markov equivalence, essential graphs and factorization. We thereafter introduce a structure learning algorithm. In the experiments, we compare the performance of our algorithm with other state-of-the-art methods on synthetic datasets. We also apply our model to investigate the protein signalling pathways and HIV-1 mutation pathways using three biological datasets.

7.3 Definitions and properties

Most notations in this section have been adapted from [13]. For a graph $G = (V, E)$ we denote with $V(G)$ the set of vertices of G and with $E(G) \subseteq V \times V$ the set of edges of G, where $E \subseteq \{\alpha \; op \; y \mid \alpha, \beta \in V \wedge op \in \{\leftarrow, \rightarrow, \leftrightarrow, -\}\}$. The symbols $\alpha \leftarrow \beta$, $\alpha - \beta$ and $\alpha \leftrightarrow \beta$ denote the directed, undirected and bidirected edges between vertices α and β, respectively. $G^U = (V, E^U)$ represents the undirected version of G called *skeleton*. The endpoint $>$ of an edge is called an *arrowhead*, or the endpoint $-$ is a *tail*. The symbol $*$ is used if the endpoint of an edge is either an arrowhead or a tail. For instance, $\alpha - *\beta$ means either $\alpha \rightarrow \beta$ or $\alpha - \beta$, and $\alpha* \rightarrow \beta$ means either $\alpha \rightarrow \beta$ or $\alpha \leftrightarrow \beta$. The *parent* set of a vertex α is $Pa(\alpha) \equiv \{\beta \mid \beta \rightarrow \alpha\}$; the *neighbor* set is $Ne(\alpha) \equiv \{\beta \mid \beta - \alpha\}$; the *spouse* set is $Sp(\alpha) \equiv \{\beta \mid \beta \leftrightarrow \alpha\}$; the *descendant* set is $De(\alpha) \equiv \{\beta \mid \alpha \rightarrow \cdots \rightarrow \beta$ or $\beta = \alpha\}$; the *ancestor* set is $An(\alpha) \equiv \{\beta \mid \beta \rightarrow \cdots \rightarrow \alpha$ or $\beta = \alpha\}$; the *anterior* set is $Ant(\alpha) \equiv \{\beta \mid \beta - * \cdots - *\alpha$ or $\beta = \alpha\}$; the *archaic* set is $Ar(\alpha) \equiv \{\beta \mid \beta* \rightarrow \cdots * \rightarrow \alpha$ or $\beta = \alpha\}$. A *path* $\pi_{\alpha,\beta}$ refers to a sequence of edges from α to β without duplicate edges.

The $\pi_{\alpha,\beta}$ is a *directed*, *undirected* or *bidirected* path if it only contains directed, undirected, or bidirected edges, respectively. A path $\alpha * - * \beta * - * \gamma$ is called a *triple* if α and γ are disjointed. A vertex β is called a *collider* on a path if and only if the path contains a triple $\alpha* \rightarrow \beta \leftarrow *\gamma$, so called *$\nu$-structure*. The $\pi_{\alpha,\beta}$ is an *inducing path* if its internal vertices are all colliders and ancestors of either α or β, or both (see examples and properties of inducing paths on the page 2815 of [13]). An undirected subgraph comprises vertices linked by undirected paths alone, whereas a bidirected subgraph comprises vertices linked by bidirected paths alone.

Given the polytree graph in Figure 1(b), we provide several examples of above definitions: $Pa(X_5) = \{X_6\}$; $Ne(X_2) = \{X_1\}$; $Sp(X_4) = \{X_5\}$; $De(X_3) = \{X_3, X_4, X_9\}$; $An(X_4) = \{X_2, X_3, X_4, X_8, X_{10}\}$; $Ant(X_4) = \{X_1, X_2, X_3, X_4, X_8, X_{10}, X_{11}, X_{15}\}$; $Ar(X_4) = \{X_2, X_3, X_4, X_5, X_6, X_8, X_{10}\}$; $X_2 \rightarrow X_3 \leftarrow X_{10}$ and $X_3 \rightarrow X_4 \leftrightarrow X_5$ are ν-structures; X_3, X_4, X_{17}, X_{20} are colliders; $\pi_{X_{10},X_{16}}$ is an undirected path; π_{X_{10},X_9} is a directed path; $\pi_{X_{17},X_{21}}$ is a bidirected path; an undirected subgraph contains X_1, X_2 and a bidirected subgraph contains X_{17}, X_{20}, X_{21}.

Let X, Y, Z be variables or sets of variables, $\langle X, Y|Z\rangle$ denotes that X and Y are conditionally independent given Z; otherwise, $\langle X, Y \nmid Z\rangle$. $\langle X, Y\rangle$ refers to the fact that X and Y are marginally independent; otherwise, $\langle X, Y \nmid \emptyset\rangle$. We use *CI* tests to determine the conditional independence. In ancestral graphs, independency relations can be identified using m-separation [14].

Definition 1. ***m-separation***. *Two vertices μ and ν are m-separated given Z in G — denoted as $\langle \mu, \nu|Z\rangle_m$ — where $Z \subseteq V(G)\backslash\{\mu, \nu\}$ if and only if every path between μ and ν contains either one triple from (1) $\alpha \rightarrow \beta \rightarrow \gamma$, $\alpha \leftrightarrow \beta \rightarrow \gamma$, $\alpha \leftarrow \beta \rightarrow \gamma$, $\alpha - \beta \rightarrow \gamma$, $\alpha - \beta - \gamma$, and $\beta \in Z$, or one triple from (2) $\alpha \rightarrow \beta \leftarrow \gamma$, $\alpha \leftrightarrow \beta \leftarrow \gamma$, $\alpha \leftrightarrow \beta \leftrightarrow \gamma$, and $De(\beta) \cap Z = \emptyset$.*

Two vertex sets X, Y are m-separated given Z if all the paths from X to Y are m-separated by Z. Figure 1(b) indicates the examples that $\langle X_2, X_4|X_3\rangle$, $\langle X_2, X_4 \nmid \emptyset\rangle$, $\langle X_2, X_8\rangle$, $\langle X_2, X_8 \nmid X_3\rangle$, $\langle X_3, X_5\rangle$, $\langle X_3, X_5 \nmid X_4\rangle$, $\langle X_2, X_9|X_3, X_4\rangle$ and $\langle X_3, X_6 \nmid X_4, X_5\rangle$.

Definition 2. ***Ancestral graph (AG)*** *[13]. A graph G is an ancestral graph if and only if three conditions hold: (i) there are no directed cycles; (ii) if there is an undirected edge $\alpha - \beta$, then α and β have neither spouses nor parents; (iii) wherever there is a bidirected edge $\alpha \leftrightarrow \beta$, no directed path passes from α to β, or from β to α.*

The attributes and examples of AG, as well as the difference between AGs and Bayesian networks, have been clarified in [13, 14]

Definition 3. ***Maximal ancestral graph (MAG)*** *[13]. An ancestral graph is maximal if and only if there exists Z such that $\langle \alpha, \beta | Z \rangle$ for any un-adjacent pair $\alpha, \beta \in V(G)$, where $Z \subseteq V(G) \backslash \{\alpha, \beta\}$.*

In ancestral graphs, vertices represent observed variables and edges represent causal relations. Many examples of AGs and MAGs were provided in [13, 15]. The polytree network in Figure 1(b) is an AG. Regarding the interpretation of directed, undirected and bidirected edges: (1) $\alpha \rightarrow \beta$ indicates that the appearance of α cultivates β which might be due to a direct cause [4]; (2) $\alpha \leftrightarrow \beta$ indicates that an unobserved variable L exists in the path $\alpha \leftarrow L \rightarrow \beta$, whereas neither α causes β nor β causes α [15]; (3) $\alpha - \beta$ indicates that α is associated with β with no certainty whether α causes β or vice-versa, due to selection bias [15].

7.4 Ancestral polytree models

In this section, we present our ancestral polytree models, which combine the concept of ancestral graphs with the idea of polytree structures.

Definition 4. ***Ancestral polytree (AP)****. A graphical model $G(V, E)$ with $E \subseteq \{\alpha \, op \, \beta \mid \alpha, \beta \in V \wedge op \in \{\leftarrow, \rightarrow, \leftrightarrow, -\}\}$ is an ancestral polytree if and only if two conditions hold: (i) it is singly connected; (ii) if it has an undirected edge $\alpha - \beta$, neither α nor β has any spouse or parent.*

Figure 1(b) demonstrates an example of AP. Moreover, it has been proven that an AG is maximal if and only if there is no inducing path between any non-adjacent vertices (Corollary 4.4, [14]). Since APs are singly connected, the conditions (i) and (ii) in the definition of AGs are guaranteed so that any AP is also an AG. Because there is no inducing path in singly connected AP so that any AP is an MAG.

As the subgraphs of causal polytrees, *causal basins* start with ν-structures, and continue in the direction of directed paths to traverse the children's descendants and the direct parents of these descendants [4]. For instance, the vertices $\{X_2, X_3, X_4, X_8, X_9, X_{10}\}$ in Figure 1(b) form a causal basin (also see examples on page 393 of [4]). We herein introduce ancestral basins in ancestral polytrees.

Definition 5. ***Ancestral basin***. *A subgraph of an ancestral polytree G is an ancestral basin G^B if it starts with a ν-structure containing a starting collider, and continues in the direction of directed or bidirected paths to pass every linked vertex, whose archaics include at least one collider, or being a parent of a collider.*

Definition 6. ***Simple ancestral polytree (SAP)***. *Ancestral polytree G is a SAP if and only if its edges are either in ancestral basins or in undirected paths.*

Proposition 1. *Both β, γ are colliders if $\beta \leftrightarrow \gamma \in E(G^S)$.*

Proof. $Ar(\beta)$ contains at least one collider since G^S is a SAP. If $Ar(\beta) \subset \{\beta\}$, then β itself is a collider; if $Ar(\beta) \supset \{\beta\}$, at least one vertex $\alpha \in Ar(\beta)$ satisfies $\alpha* \rightarrow \beta \leftrightarrow \gamma$. It ensures that β is a collider, so is γ. □

A *starting collider*, also termed as a multi-parent node [4] or articulation point [6], represents the vertex β on a path containing $-\alpha \rightarrow \beta \leftarrow \gamma-$, where an ancestral basin begins. For instance, the subgraph containing $\{X_2, X_3, X_4, X_5, X_6, X_8, X_9, X_{10}\}$ in Figure 1(b) is an ancestral basin whose starting collider is X_3. Yet, neither the subgraph containing $\{X_{11}, X_{12}, X_{13}, X_{14}\}$ nor the one containing $\{X_{15}, X_{17}, X_{18}, X_{19}, X_{20}, X_{21}\}$ is an ancestral basin, because there is no starting collider in both subgraphs, and neither the parents of X_{11}, X_{19} nor the anteriors of X_{11}, X_{19} contain any collider. The AP in Figure 1(b) becomes an SAP after replacing $X_{12} \rightarrow X_{13}$, $X_{19} \rightarrow X_{18}$ with $X_{12} \leftarrow X_{13}$, $X_{19} - X_{18}$, respectively.

Let $H(G)$ represent the independency in a causal diagram:

$$H(G) \equiv \{\langle X, Y|Z\rangle \mid \textit{Disjointed subsets } X, Y, Z \subseteq V(G)\}$$

Definition 7. ***Markov equivalence*** *[13]. Two ancestral graphs G_1 and G_2 are Markov equivalent, $G_1 \sim G_2$, if $H(G_1) = H(G_2)$.*

Proposition 2. *$G_1 \sim G_2$ if and only if APs G_1 and G_2 have the same skeletons and ν-structures.*

Due to limited space, a concise proof includes two parts. (Necessary) Note that both APs G_1 and G_2 are also MAGs. If G_1, G_2 are MAGs and $G_1 \sim G_2$, then G_1, G_2 have the same adjacencies and ν-structures (Proposition 3.6 in [13]). (Sufficient) It was proven that if MAGs G_1, G_2 share the same skeleton and colliders with order, then $G_1 \sim G_2$ (Theorem

3.7, [13]). If APs G_1 and G_2 have the same ν-structures, they have the same colliders with order because there is no discriminating path (definition 3.8 and 3.11, [13]) in any AP.

Above proposition leads to a sufficient and necessary condition to identify Markov equivalent APs. However, instead of investigating all equivalent structures, it is vital to identify one essential graph which is sufficient to represent all Markov equivalent structures.

Definition 8. ***Essential graph***. *Let $[G]$ denote Markov equivalence class of ancestral polytree G, whose conditional independencies are identical. M_G is the essential graph of G if and only if M_G satisfies two conditions, (i) M_G shares the same skeleton with all APs in $[G]$; (ii) any directed or bidirected edge exits in M_G if and only if it is shared by all APs in $[G]$.*

Note that essential graphs are equivalent to partial ancestral graphs [15] if AGs are restricted to be polytrees. Let G^S be a SAP, due to the condition (ii) above, $G^S \in [G]$ guarantees that G^S includes all directed and bidirected edges in M_G. This fact leads to the next proposition.

Proposition 3. *M_G is a simple ancestral polytree.*

The factorization of acyclic directed mixed graphs with directed and bidirected edges was studied in [16]. The factorization of Gaussian distribution in MAG can be decomposed as $f(X_V) = f(X_{un_G})f(X_{V\setminus un_G} \mid X_{un_G})$, where un_G and $V\setminus un_G$ contain vertices of undirected and directed subgraphs in MAG, respectively [14]. To assess structure learning and inference, the next proposition clarifies the factorization of joint probability of ancestral polytrees.

Proposition 4. *Given an ancestral polytree $G = (V, E)$, the joint probability of random variables can be decomposed into products of conditional probability distributions as:*

$$P_G(X_V) = \frac{1}{Z} \prod_{X_i - X_j \in E(S^U)} \psi(X_i,X_j)\times\prod_{b\in S^B} P(X_b|Ant(X_b))\times\prod_{X\in S^D} P(X|Ant(X),Ar(X))$$

Where S^B is the set of subgraphs containing bidirected edges, S^U is the set of subgraphs containing undirected edges, S^D is the set of subgraphs containing directed edges, ψ is a factor potential of a clique formed with edge $\alpha - \beta$ as a non-negative function, the normalization coefficient is defined as $Z = \int_{X_{V(S^U)}} \prod_{\alpha-\beta\in S^U} \psi(\alpha, \beta)dX$.

Proof. Firstly, given any undirected subgraph s in S^U and any bidirected subgraph b in S^B, the factorization can be expressed as $\prod_{c\in C(s)} \psi_c(X)/Z^*$, where $C(s)$ is the set of cliques in s and Z^* is the normalization coefficient. Note that undirected subgraphs in S^U are disjointed and cliques in singly connected S^U contain two neighbouring nodes α, β on an undirected edge $\alpha - \beta$. Therefore, the factorization of S^U is:

$$P(X_{S^U}) = \prod_{s\in S^U} \frac{1}{Z^*} \prod_{c\in C(s)} \psi_c(X) = \frac{1}{Z} \prod_{c\in C(S^U)} \psi_c(X) = \frac{1}{Z} \prod_{X_i - X_j \in E(S^U)} \psi(X_i, X_j)$$

Secondly, Theorem 4 in [16] reveals the factorization of bidirected subgraphs as:

$$P(X_{S^B}|Ne(X_{S^B}), X_{S^U}) = \prod_{b\in S^B} P(X_b \mid Pa(X_b), V(S^U)) = \prod_{b\in \mathcal{S}^B} P(X_b \mid Ant(X_b))$$

Thirdly, the factorization of directed subgraphs is:

$$P(X_{S^D}|Ne(X_{S^D}), X_{S^U}) = \prod_{X\in S^D} P(X \mid Pa(X), V(S^B), V(S^U)) = \prod_{X\in S^D} P(X \mid Ant(X), Ar(X))$$

The proof is complete by multiplying three parts:

$$\begin{aligned} P_G(X) &= P(X_{S^U})P(X_{S^B}, X_{S^D}|X_{S^U}) \\ &= P(X_{S^U})P(X_{S^B}|Ne(X_{S^B}), X_{S^U})P(X_{S^D}|Ne(X_{S^D}), X_{S^U}) \end{aligned}$$

□

Particularly, if the entire ancestral polytree G is an ancestral basin, we have $S^U = \emptyset, S^D = V - V(S^B), Ant(X) = Ar(X)$. Mimicking each vertex in S^D as an entire bidirected subgraph, the factorization of ancestral basin G^B is:

$$\begin{aligned} P_{G^B}(X_V) &= \prod_{X\in V - V(S^B)} P(X \mid Ant(X)) \times \prod_{b\in S^B} P(X_b \mid Ant(X_b)) \\ &= \prod_{b\in S^B \cup (V - V(S^B))} P(X_b \mid Ant(X_b)) \end{aligned}$$

Examples in Figure 2 include: $V(S^U) = \{\{X_1, X_2\}, \{X_6, X_7\}, \{X_{10}, X_{11}, X_{15}, X_{16}\}\}, V(S^B) = \{\{X_4, X_5\}, \{X_{17}, X_{20}, X_{21}\}\}, V(S^D) = \{\{X_9\}, \{X_{12}, X_{13}, X_{14}\}, \{X_3, X_8\}, \{X_{18}, X_{19}\}\}$.

The factorization of AP with configuration $X = x$ in Figure 2 is: $P(X = x) = [1/Z$ $\psi(x_1, x_2)\psi(x_6, x_7)$ $\psi(x_{10}, x_{11})\psi(x_{11}, x_{15})\psi(x_{15}, x_{16})] \times [p(x_{19})p(x_{18}|x_{19})p(x_{13}|x_{12})p(x_{14}|x_{12})$ $p(x_3|x_1, x_2, x_8, x_9, x_{10}, x_{11}, x_{15}, x_{16})p(x_{12}|x_{10}, x_{11}, x_{15}, x_{16})] \times [p(x_4, x_5|x_3, x_6, x_7)$ $p(x_{17}, x_{20}, x_{21}|x_{10}, x_{11}, x_{15}, x_{16}, x_{18})]$

7.5 Learning ancestral polytrees

Given a training dataset, the methodology of training causal polytrees usually involves two stages. Firstly, undirected skeletons are trained either by maximum weighted spanning trees [4] or by CI tests [2]. Secondly, the directionality of edges in undirected skeletons are recovered by orienting principles [4, 6] including two rules. Rule 1: for all $\alpha - \beta - \gamma$ and $\langle \alpha, \gamma \rangle$, orient $\alpha - \beta - \gamma$ into $\alpha \rightarrow \beta \leftarrow \gamma$. Rule 2: for remaining $\alpha \rightarrow \beta - \gamma$, orient $\alpha \rightarrow \beta - \gamma$ into $\alpha \rightarrow \beta \rightarrow \gamma$. Based on the m-separation, orienting principles for learning ancestral polytrees rely on the orienting principles of ancestral polytree (OPAP), which also includes two rules:

Rule 1 : for all $\alpha * - * \beta * - * \gamma$ and $\langle \alpha, \gamma \rangle$, orient $\alpha * - * \beta * - * \gamma$ into $\alpha * \rightarrow \beta \leftarrow * \gamma$.

Rule 2 : for remaining $\alpha * \rightarrow \beta - \gamma$, orient $\alpha * \rightarrow \beta - \gamma$ into $\alpha * \rightarrow \beta \rightarrow \gamma$.

Based on the OPAP, we have designed a structure learning algorithm using three tips: (1) search colliders through visiting inner vertices from the one with the most undirected edges to the one with the least, because every collider is an inner vertex. (2) For each selected inner vertex, examine CI test $\langle \alpha, \gamma \rangle$ on the triplet $\alpha - \beta - \gamma$ first, and examine CI test on $\alpha \leftarrow \beta \rightarrow \gamma$ last (because it is comparatively rare for both α and γ to be colliders). (3) Distinguish bidirected from undirected edges by withdrawing detected bidirected edges, and recover them back into the oriented structure.

Algorithm: Learning Ancestral Polytree (LAP)

Input: A training dataset and a polytree skeleton G^U.

Output: A partially oriented ancestral polytree G.

Abbreviations: *UV*, the set of unvisited inner vertices; *CS*, the set of colliders; *BA*, the set of bidirected edges; V_{In}, the set of inner nodes; *CI*, the set of CI test.

Initiate $CI = BA = \emptyset$, $UV = V_{In}, G = G^U$.

While $UV \neq \emptyset$

$\beta = \arg\max_{v \in UV} |Ne_G(v)|$; $UV = UV \backslash \{\beta\}$;

For all $\alpha - \beta - \gamma \in E(G)$ and $(\alpha, \gamma) \notin CI$

Do $CI = CI \cup \{(\alpha, \gamma)\}$; if $\langle \alpha, \gamma \rangle$, orient $\alpha - \beta - \gamma$ into $\alpha \rightarrow \beta \leftarrow \gamma$ and $CS = CS \cup \{\beta\}$;

End for

For all $\alpha - \beta \leftarrow \gamma \in E(G)$ and $(\alpha, \gamma) \notin CI$

Do $CI = CI \cup \{(\alpha, \gamma)\}$; if $\langle \alpha, \gamma \rangle$, orient $\alpha - \beta \leftarrow \gamma$ into $\alpha \rightarrow \beta \leftarrow \gamma$ and $CS = CS \cup \{\beta\}$;

End for

For all $\alpha - \beta \rightarrow \gamma \in E(G)$ and $(\alpha, \gamma) \notin CI$

Do $CI = CI \cup \{(\alpha, \gamma)\}$; if $\langle \alpha, \gamma \rangle$, orient $\alpha - \beta \rightarrow \gamma$ into $\alpha \rightarrow \beta \leftrightarrow \gamma$ and $CS = CS \cup \{\beta\}$,$E(G) = E(G) \backslash \{\beta \leftrightarrow \gamma\}, BA = BA \cup \{(\beta, \gamma)\}$;

End for

For all $\alpha \leftarrow \beta \rightarrow \gamma \in E(G)$ and $(\alpha, \gamma) \notin CI$

Do $CI = CI \cup \{(\alpha, \gamma)\}$; if $\langle \alpha, \gamma \rangle$, orient $\alpha \leftarrow \beta \rightarrow \gamma$ into $\alpha \leftrightarrow \beta \leftrightarrow \gamma$ and $CS = CS \cup \{\beta\}$, $E(G) = E(G) \backslash \{\alpha \leftrightarrow \beta, \beta \leftrightarrow \gamma\}, BA = BA \cup \{(\alpha, \beta), (\beta, \gamma)\}$;

End for

For all $\alpha \rightarrow \beta - \gamma \in E(G)$

Do orient $\alpha \rightarrow \beta \rightarrow \gamma$;

End for

End while

For all $\alpha \in CS$, create an empty queue Q, $push(Q, \alpha)$;

While $Q \neq \emptyset$

$\beta = pop(Q), UV = UV \backslash \{\beta\}$;

For all $\gamma \in UV, \beta - *\gamma \in E(G)$

Do orient $\beta - \gamma$ into $\beta \leftarrow \gamma$, $push(Q, \gamma)$;

End for

End while

End for

For all $(\alpha, \beta) \in BA$

Do $E(G) = E(G) \cup \{\alpha \leftrightarrow \beta\}$.

End for

Return G

LAP has two major parts. The first part in the "while" loop and the second part in the "for" loop orient the unvisited edges based on the rule 1 and rule 2 of OPAP, respectively. Using CI tests, each of the five "for" loops in the first part orients $\alpha - \beta - \gamma, \alpha - \beta \leftarrow \gamma, \alpha - \beta \rightarrow \gamma, \alpha \leftarrow \beta \rightarrow \gamma$ and $\alpha \rightarrow \beta - \gamma$, respectively. The visited edges and nodes are recorded to avoid repeating tests. The second part uses the depth-first traversal to visit all colliders in ancestral basins and to orient the undirected edges subsequently.

Ideally, consider there is an oracle to provide CI information from a faithful ancestral polytree G, denoted as $G_{\mathcal{T}}$. An algorithm is *sound* if it outputs a predicted G that $G \in [G_{\mathcal{T}}]$, and algorithm is *complete* if it predicts a maximally informative G for $[G_{\mathcal{T}}]$ [15]. The soundness and completeness of 11 orientation rules have been proven to train ancestral graphs [15]. These 11 orientation rules can be simplified into OPAP if AGs are singly connected, which ensures the soundness and completeness of LAP. Remind that two OPAP rules were also included in the Fast Causal Inference algorithm (*FCI*) [17] and the augmented *FCI* [15], both of which can model equivalent structures as LAP if underlying structures were singly connected.

Many studies have endeavored to learn polytree skeletons and well-known algorithms for maximum-likelihood learning of tree distributions have achieved the complexity of $\mathcal{O}(n^2 \log(n))$ [18]. Herein we analyze the computational complexity of LAP to show that LAP can achieve a fast structure learning using refined orienting procedures.

Proposition 5. *Suppose skeleton G^U is an undirected tree which has one root with K adjacent vertices, and has inner vertices all with K+1 adjacent vertices. Let N be the number of vertices in G^U, the number of required CI tests $R(G^U)$ satisfies:*

$$R(G^U) \leq (K+1)(N-1)/2 - K$$

Proof. Let H be the depth of tree G^U, we have $N = \sum_{i=1}^{H} K^{i-1} = (K^H - 1)/(K-1)$. Therefore, $\sum_{i=2}^{H-1} K^{i-1} = (K^{H-1} - K)/(K-1) = (N-1)/K - 1$. The number of required CI tests is maximal if we test $\langle \alpha, \gamma \rangle$ in all triplets formed as $\alpha - \beta - \gamma$. Consider every inner vertex has $K(K+1)/2$ pairs of adjacent vertices except the root, the summation of CI tests over all inner vertices satisfies:
$R(G^U) \leq K(K-1)/2 + K(K+1)\sum_{i=2}^{H-1} K^{i-1}/2 = (K+1)(N-1)/2 - K.$ □

Proposition 5 analyzes the complexity of LAP for a special case of skeleton G^U. For the general cases, we then consider the average CI tests regarding the entire space of the set of

marginally independent tests $\mathcal{T}$, denoted as Ω. Given an arbitrary G^U, let $f(n)$ count the number of $\mathcal{T}$ that requires n CI tests to train G^U. In fact, the set of $f(n)$ has $k(k-1)/2 - \lceil k/2 \rceil + 1$ elements where $n = \lceil k/2 \rceil, \lceil k/2 \rceil + 1, \cdots, k(k-1)/2$ and $\lceil k/2 \rceil$ denotes the upper bound of integer upon $k/2$ (e.g. $\lceil 3/2 \rceil = 2$). The average CI tests of inner node v is:

$$E[R(G_v^U)] = \frac{1}{|\Omega|} \sum_{n=\lceil k/2 \rceil}^{k(k-1)/2} [n \times f(n)]$$

To date, we have not found any simple formula to decode $f(n)$ whose experiment data is: f(2)={1},f(3)={2,6},f(4)={16,8,8,14,18},f(5)={128,192, 192,224,104,72,62,50}, f(6)={4096, 4096,4096,3584,3584, 3968,3520,2560,1776,744,392,222,130}, f(7)={131072,262144, 327680, 311296,262144,233472,169984,123904,88064,66560, 50048,33856,19808,10480,3944,1672, 702,322}, f(8)={16777216,25165824, 29360128,29360128,27787264,25165824, 22413312, 18874368,14680064,10485760,7061504,4673536, 2883584,1775616,1086464,681728,415488, 233792,116128,51248,17384,6152,2046, 770}.

Even so, we have observed that regression methods can estimate $E[R(G_v^U)]$ sufficiently. Particularly, we have found similar results using both least square and robust linear regressions based on iteratively re-weighted least squares. The output of least square regression is: $E[R(G_v^U)] = -1.3439 + 1.3425K$, where the root mean square error is $\varepsilon = 0.2652$ and the maximum residue at $K = 5$ equals to $\varepsilon_{max} = 0.3126$. Based on the above estimation, we have:

$$E[R(G^U)] = \sum_{V_i \in V_{In}} [\frac{1}{|\Omega|} \sum_{T \in \Omega} R(G_{V_i}^U, \mathcal{T})] = \sum_{V_i \in V_{In}} E[R(G_{V_i}^U)]$$

If $K = \max_{v \in V} |Ne_{G^U}(v)| \leq 8$, $E[R(G^U)] \leq |V_{In}| \times (1.3425K - 1.3439 + \varepsilon_{max})$. In other words, the average CI tests for each inner node are bound by 1.3425K-1.0313 if the maximal graphical degree meets the condition of K$\leq$8.

7.6 Experiments

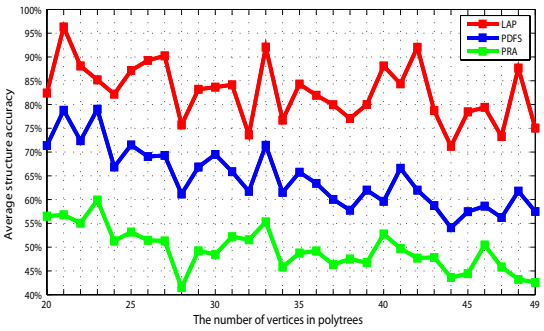
Four experiments were carried out in this study: (1) we compared LAP with both the causal Polytree Recovery Algorithm (PRA) [4] and the Polytree-Depth-First-Search (PDFS) algorithm [6] using a synthetic dataset, (2) we applied LAP to model protein signaling pathways using a human immune cell dataset [12], (3) we explored the HIV-1 resistance mutation path-

ways by LAP using the HIV-1 protease inhibitor nelfinavir (NFV) dataset [19], and (4) we used LAP to model the interaction networks in the HIV-1 capsid protein.

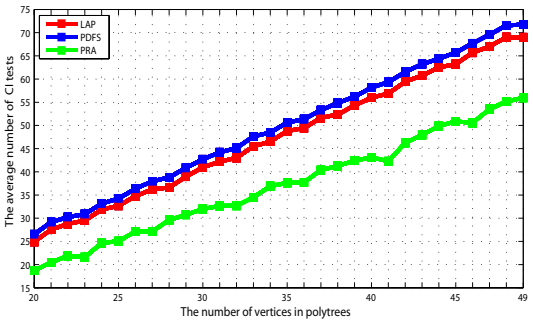
In our experiments, the algorithm performance was evaluated using structure accuracy defined as: $Acc = (|\{(X,Y)|(X,Y) \in E, (X,Y) \in E^*, (Y,X) \notin E^*\}|)/(|V|-1)$, where $G = (V, E)$ and $G^* = (V, E^*)$ are the known and the predicted structures, respectively. We used non-parametric bootstrapping with 100 replicates to cultivate polytree skeletons from the undirected maximal spanning tree algorithm based on mutual information [20]. We used Fisher's exact test for CI tests (confidence level: 95%) and Laplace smoothing for probability calculations.Our algorithm implementation is available at `http://www.mathworks.com/matlabcentral/fileexchange/40126-ancestral-polytree`.

In the first experiment, we compared LAP with two causal polytree algorithms PDFS and PRA using a synthetic dataset. The synthetic data contained 6000 random polytrees with variable numbers (ranging from 20 to 49) and CI sets randomly generated. For each variable number, 200 unique skeletons were randomly sampled for our analysis. Figure 2(a) illustrates the comparison results, showing that LAP (average structure accuracy: 82.72%) performs better than the PDFS (64.6%) and PRA (49.54%) algorithms. We also compared the number of CI tests required by the polytree algorithms, illustrated in the Figure 2(b). The average CI tests for PDFS, LAP and PRA were found to be 49.51, 47.53 and 37.02, respectively. This suggests that the weak structure accuracies of PRA are compromised by the least CI tests, and LAP performs better than PDFS.

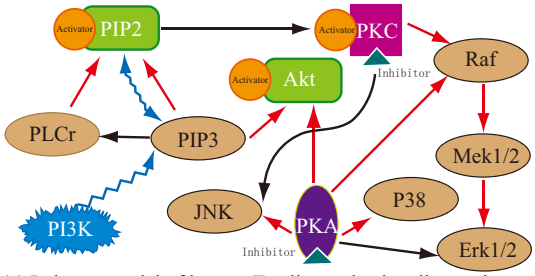
In the second experiment, LAP was applied to flow cytometry datasets of human T cell protein signaling pathways [12]. The data was collected through intracellular multicolor flow cytometry, which measures the protein expression levels of 11 proteins with single-cell data points [12]. We first removed the outliers whose values were 3 times larger than the mean values and discretized the continuous data using an information preserving algorithm [21]. By doing so, 500 datasets were created containing data for 400 cells each. Using 100 bootstrap samples on each dataset lead to 500 trained APs. Figure 2(c) shows a consensus AP which recovers 10 out of 14 expected signaling pathways, while BN analysis recovered 12 [12]. Note that the bidirected edge PIP2↔PIP3 was recovered assuming that the protein PI3K was observed.



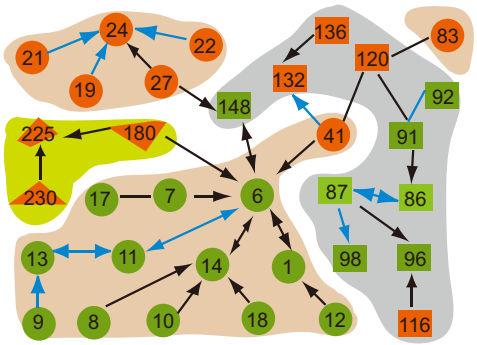
(a) Structure accuracy of learning algorithms



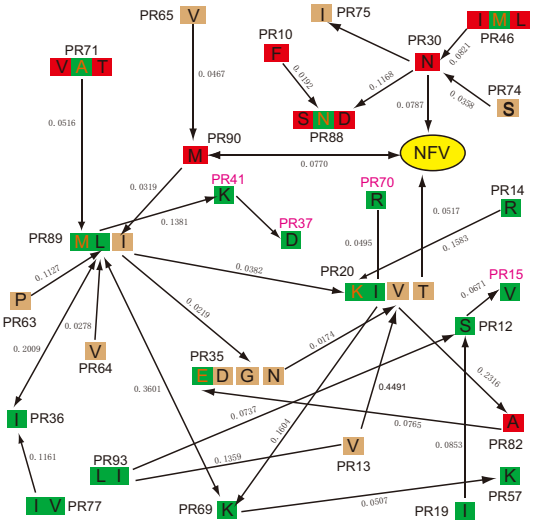
(b) Number of CT tests required by learning algorithms



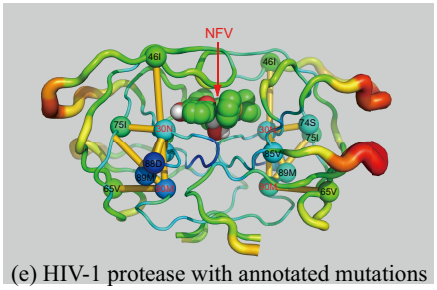
(c) Polytree model of human T cell protein signaling pathways



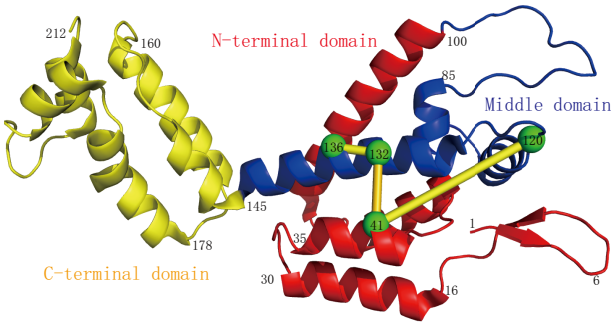
(f) Polytree model of HIV-1 capside positions



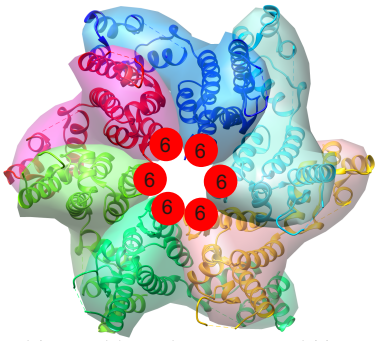
(d) Polytree model of HIV-1 protease mutation pathways



(e) HIV-1 protease with annotated mutations



(g) HIV-1 capsid with annotated positions



(h) Residue position 6 in HIV-1 capsid hexamer

Figure 2. Average structure accuracy (a) and number of CI tests (b) required by LAP, PDFS

and PRA on the a synthetic dataset.(c) Human T cell protein signaling pathway network modeled by LAP. Pink circles and blue triangles represent activators and inhibitors respectively. Other nodes represent proteins in the signaling pathways. Red arrows are faithfully predicted edges, black arrows are undetected edges but reported by biological studies and blue arrows are recovered assuming that the protein PI3K is observed [12]. (d) Ancestral polytree network of HIV-1 protease mutations selected by the protease inhibitor NFV. NFV is colored yellow and colored squares distinguish protease drug resistance mutations from wild type residues. Mutations from the same residue position are clustered and mutual information is notified on edges. (e) HIV-1 protease structure. The residue positions are annotated accordingly. The PDB data of HIV-1 protease is 2QAK and the visualization software is PyMOL v1.5. (f) Ancestral polytree modeling of residue interaction networks in HIV-1 capsid. Green and pink indicate residue positions in the loop and helix structures of capsid, respectively. Circle, square and triangle represent positions in the C-terminal (1-84), middle (85-145) and N-terminal (146-231) domains. Red edges indicate positions whose C_α atoms are closer than 10Å of the Euclidean distance in the capsid structure. (g) HIV-1 capsid structure (PDB:3P05, visualized by PyMOL v1.5). The residue positions 41, 120, 132, 136 across different functional domains are annotated, as well as major clusters in two loop regions (positions: 1-16 and 85-100). Yellow links indicate the associations between residue positions 136, 132, 41 and 120, predicted by our ancestral polytree model. (h) Position 6 in the structure of HIV-1 capsid hexamer (PDB:3H4E, visualized by Chimera v1.7).

In the third experiment, we modeled the interaction network of HIV-1 protease mutations selected by the protease inhibitor nelfinavir (NFV). We trained polytree networks using the HIV-1 NFV dataset, which includes 1307 HIV-1 protease amino acid sequences sampled from 967 drug-naive and 340 NFV-treated patients [19]. The trained polytree network is illustrated in Figure 2(d). We mapped the mutation positions to the crystalized structure of HIV-1 protease (Figure 2(e)). We found that for protease mutations that had less than 5 edges to the NFV in the consensus AP, the Euclidean distances of C_α atoms between these mutations had less than 10 angstroms in the 3D structure. Moreover, our AP shared 38 out of 58 edges compared to the trained BNs, described previously in [19]. AP may help to study the associations between HIV-1 drug resistance mutations, whereas causal effects in drug resistance pathways and the impact of unobserved variables require further investigations.

In our last experiment, we modeled the interaction networks of natural residues in HIV-1 capsid - a hexamer protein in the length of 231 residue positions. HIV-1 capsid is a key protein to construct the structural surface of HIV-1 mature virions [22]. We followed the procedure described in [22] to collect HIV-1 capsid sequences sampled from 787 treatment naive patients in the HIV Los Alamos database. We removed both duplicate sequences and sequences with stop codons. We aligned nucleotide sequences using Seaview v4.4.0. Figure 2(f) shows our consensus ancestral polytree for HIV-1 capsid. It suggests that positions are clustered in AP regarding to the loop and helix regions in functional domains of capsid (Figure 2(g)), except the positions 83 and 116. Being crucial for protein multimerization, position 6 connected with many positions in AP (Figure 2(h)). Potentially, positions from the same functional domains tempt to cluster in our causal network may explain the role of structural constraints. As shown in our recent study [22], this information can be useful for designing novel inhibitors targeting HIV-1 capsid.

7.7 Conclusions and future work

This study introduces ancestral polytrees and their simple structures which guarantee fast learning algorithms. Our future study will focus on maximum likelihood and inference problems. We will also apply polytree models to large biological networks such as genomic interaction networks.

References

1. Pearl J (2000) Causality: models, reasoning and inference. Cambridge Univ Press.
2. Huete JF, de Campos LM (1993) Learning causal polytrees. In: Symbolic and Quantitative Approaches to Reasoning and Uncertainty. Springer, pp. 180–185.
3. Rebane G, Pearl J (1987) The recovery of causal poly-trees from statistical data. In: 3rd Conf. on UAI. pp. 222-228.
4. Pearl J (1988) Probabilistic Reasoning in Intelligent Systems: Networks of Plausble Inference. Morgan Kaufmann Pub.
5. Dasgupta S (1999) Learning polytrees. In: 15th Conf. on UAI. pp. 134–141.
6. Ouerd M, Oommen B, Matwin S (2004) A formal approach to using data distributions for building causal polytree structures. Information Sciences 168: 111–132.

7. Campos LMd (1998) Independency relationships and learning algorithms for singly connected networks. Journal of Experimental & Theoretical Artificial Intelligence 10: 511–549.

8. Pena JM (2007) Reading dependencies from polytree-like bayesian networks. In: 23th Conf. on UAI. pp. 303–309.

9. Lodish H, Berk A, Matsudaira P, Kaiser CA, Krieger M, et al. (2004) Molecular cell biology. WH Freeman, 5th edition.

10. Messaouda O, Oommen JB, Matwin S (2003) Enhancing caching in distributed databases using intelligent polytree representations. In: Advances in Artificial Intelligence. Springer, pp. 498–504.

11. Zaveri MS, Hammerstrom D (2010) Cmol/cmos implementations of bayesian polytree inference: Digital and mixed-signal architectures and performance/price. Nanotechnology, IEEE Transactions on 9: 194–211.

12. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP (2005) Causal protein-signaling networks derived from multiparameter single-cell data. Science 308: 523.

13. Ali RA, Richardson TS, Spirtes P (2009) Markov equivalence for ancestral graphs. The Annals of Statistics 37: 2808–2837.

14. Richardson T, Spirtes P (2002) Ancestral graph markov models. The Annals of Statistics 30: 962–1030.

15. Zhang J (2008) On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. Artificial Intelligence 172: 1873–1896.

16. Richardson TS (2009) A factorization criterion for acyclic directed mixed graphs. In: 25th Conf. on UAI. pp. 462–470.

17. Spirtes P, Meek C, Richardson T (1995) Causal inference in the presence of latent variables and selection bias. In: 11th Conf. on UAI. pp. 499–506.

18. Leiserson CE, Rivest RL, Stein C, Cormen TH (2001) Introduction to algorithms. The MIT press.

19. Deforche K, Silander T, Camacho R, Grossman Z, Soares M, et al. (2006) Analysis of hiv-1 pol sequences using bayesian networks: implications for drug resistance. Bioinformatics 22: 2975–2979.

20. Chow C, Liu C (1968) Approximating discrete probability distributions with dependence trees. Information Theory, IEEE Transactions on 14: 462–467.

21. Hartemink AJ (2001) Principled computational methods for the validation of and discovery of genetic regulatory networks. Ph.D. thesis, MIT.

22. Li G, Verheyen J, Rhee SY, Voet A, Vandamme AM, et al. (2013) Functional conservation of hiv-1 gag: implications for rational drug design. Retrovirology 10: 126.

Chapter 8

General discussion and future perspectives

"If I have seen further it is by standing on the shoulders of giants."

— Isaac Newton

The contribution of my Phd thesis can be summarized in one sentence: it contributes to the understanding of HIV genome-wide diversity, interaction and coevolution. My research has been involved mainly with two topics: HIV genomic analysis and the development of computational methods. HIV genomic projects have been designed toward clinical applications (e.g. HIV drug resistance) as well as drug and vaccine development. I have developed computational methods for large-scale data analysis mostly based on the mathematical modeling. My software toolboxes have been shared as open sources in the journal websites and in the Matlab file exchange center (ID: 45984).

In the last chapter of my thesis, I will summarize the contributions of my Phd study and provide original ideas for future research. The strength and weakness of my HIV projects are also discussed. At the end of this chapter, my contributions on probabilistic graphical models and data visualization are briefly discussed as two side projects of my Phd study. Future perspectives of my genome-wide analysis on HBV and HCV are discussed as well.

8.1 HIV-1 genetic diversity and drug resistance

In the past three decades, thousands of anti-HIV inhibitors have been proposed and more than 25 inhibitors in five drug classes have been approved by the FDA (Chapter 1). Nevertheless, many HIV clinical studies have demonstrated that the efficiency of anti-HIV inhibitors have been hampered by drug resistance mutations emerging in the HIV genome [1]. Usually, anti-HIV drugs with low genetic barriers have higher tendency towards development of drug resistance, while those with higher genetic barriers have a lower tendency towards development of drug resistance [2]. All NNRTIs and most NRTIs are considered to have lower genetic barriers than protease inhibitors [2]. Previous HIV-1 studies have investigated the role of protease mutations in the drug resistance of protease inhibitors, which have relatively higher drug genetic barriers. As a new mechanism of drug resistance, HIV-1 Gag mutations have been found to cause drug resistance to protease inhibitors [3]. The presence of Gag mutations has been associated with protease drug resistance [4]. Nevertheless, large-scale analysis on PI-associated Gag mutations is largely lacking.

Many HIV drug resistance interpretation algorithms have been designed to estimate genotypic drug resistance using viral sequences. These algorithms are built on the drug resistance mutations reported in experimental and clinical studies. Currently, several drug resistance interpretation algorithms are widely used; for instance, the Stanford algorithm (http://sierra2.stanford.edu/sierra/servlet/JSierra), the Rega algorithm (https://rega.kuleuven.be/cev/avd/software/rega-algorithm), the ANRS algorithm (http://www.hivfrenchresistance.org/) and the drug resistance mutation list from International Antiviral Socienty-USA (https://www.iasusa.org/). HIV clinical guidelines from WHO and NIH have also been updated regularly (http://aidsinfo.nih.gov/). Overall, advancements on many aspects of detecting and monitoring drug resistance have been made. Improved HIV treatments have further prolonged the life expectancy of HIV infected patients [5].

Ideas of my HIV-1 diversity and drug resistance studies

HIV *gag* gene encodes important structural proteins for viral morphogenesis (see Chapter 1). Gag polyproteins are cleaved into individual structural proteins during the protease-mediated proteolytic processing. Genetic diversity of Gag proteins can impact on the synthesis and maturation of viral structural proteins [6]. In my thesis,

Chapter 2 reports the genetic diversity of HIV-1 Gag and Chapter 4 investigates HIV-1 Gag mutations associated with drug resistance of protease inhibitors.

During the process of data collection and literature review, I found that many experimental inhibitors had been designed to target Gag but few studies had reported the genetic diversity of drug binding sites in the HIV-1 Gag. Since natural polymorphisms can affect the potential of experimental Gag inhibitors [7], I conducted large-scale sequence analysis to reveal the functional conservation of HIV-1 Gag and to show all possible natural polymorphisms observed at the known drug binding sites (Chapter 2).

In the meantime, Dr. Kristof Theys and Prof. Kristel Van Laethem helped to contact Dr. Jens Verheyen whose expertise is known in the field of HIV Gag mutations. After a few months, HIV-1 *gag* nucleotide sequences and clinical data from our Leuven patient cohort were available for me to identify HIV-1 Gag mutations emerging during the treatment of protease inhibitors. With the help of my colleagues, I drafted the manuscript about the drug resistance of HIV-1 Gag mutations (Chapter 4).

Strength and weakness of my HIV-1 Gag studies

Chapter 2 and 4 characterize genetic variations of HIV-1 Gag and its impact on the protease drug resistance. In these two studies, large-scale datasets were used in our statistical analysis. Previous studies have reported results only in small cohorts of patients, whereas I analyzed the data sampled from more than 10000 patients. Chapter 2 provides the first study to show the functional conservation of HIV-1 Gag proteins. We showed natural variations at drug binding sites of over 50 Gag experimental inhibitors and revealed the N terminus of capsid as the most conserved drug target in HIV-1 Gag.

Chapter 4 shows for the first time the prevalence of PI-associated Gag mutations in large-scale cohorts of patients infected with different HIV-1 subtypes. We showed that most Gag mutations were less prevalent than previously thought. Only a few significant Gag mutations were found in the C terminus of HIV-1 Gag. Our study thus contributes to the prevalence analysis of HIV-1 PI-associated Gag mutations in large-scale patient populations.

The progress of my HIV-1 Gag projects has faced many challenges. I was unable to further investigate and quantify the clinical impact of point mutations in HIV-1 Gag.

Owing to the limited resources, I was unable to perform the *in vitro* experiments to quantify the drug resistance levels of Gag mutations. Future perspectives of these two studies are associated with the development of new Gag inhibitors (Chapter 2) and the evaluation of clinical impact of PI-associated Gag mutations (Chapter 4). The former will take years to develop new HIV-1 Gag inhibitors; the latter will require large clinical and experimental data which is largely lacking to date. It remains a challenge to address the virological outcome of HIV-1 Gag mutations and the development of Gag inhibitors. In addition, different HIV-2 Gag mutations have been shown to associate with PI drug resistance compared to HIV-1 [8]. Further investigations on HIV-2 Gag mutations are still needed.

8.2 HIV genomic diversity

The HIV pandemic has been characterized by extensive genomic diversity caused by multiple factors including high evolutionary rates, sequence recombination and multiple zoonotic transmissions into human populations [9]. In the multi-national vaccine trial STEP, 252 full-length genome sequences of subtype B were sampled from 42 patients to show the genetic difference between the vaccine and the placebo groups [10]. In the Thai vaccine trial RV144, 359 full-length genome sequences of CRF01_AE were extracted from 49 patients to demonstrate that vaccine-induced immune responses were associated with residues in the variable regions of GP120 [11, 12]. Moreover, HIV-1 transmission and viral evolution were characterized using 475 subtype B genome sequences sampled from 11 American patients during acute HIV-1 infection [13]. Subtype distributions in South Africa and Malaysia were assessed using 244 and 184 genome sequences, respectively [14, 15]. Investigations of HIV genome sequences sampled from less than 100 patients have also been reported [16-25]. A few studies have also quantified the HIV genome-wide diversity using small sequence datasets [16-18]. Since HIV genome-wide diversity has a significant effect on the development of HIV drugs and vaccines [9], one of our objectives was set up to investigate the genome-wide diversity of HIV from a global perspective.

Ideas of my HIV genome projects

My idea to investigate HIV genome-wide diversity was inspired by HIV-1 Gag conservation study (Chapter 2). Most results of HIV-1 genome-wide diversity analysis are presented in Chapter 3. The original idea originated from conversations

with Dr. Kristof Theys a year ago. We had read that HIV genomic diversity reported discordantly by different publications, raising our interests to publish results about HIV genome-wide diversity. When this clear goal was set up, I proceeded to collect all full-length genomic sequences from public databases. During my previous projects, I developed the toolbox for genome sequence alignment and collected many genomic datasets (e.g. human T cell epitopes, HIV-human protein interaction, HIV-derived peptide inhibitors). After the initiation of my HIV genome diversity study, I applied my alignment toolbox and integrated the genomic datasets that I collected in my previous projects.

Strength and weakness of my genomic studies

I performed comparative analyses to accurately estimate HIV genomic diversity using large-scale genomic datasets (Chapter 3). I have further contributed to an open-source genomic toolbox developed for genomic analysis. However, due to limited time and resource, I could not use the information of HIV genomic diversity to improve current inhibitors and bring much improvement in the HIV clinical treatments and vaccine trials.

Many technologies in the field of viral infectious diseases have been advanced in the past few years (e.g. high throughput genomic sequencing, database management, statistical methodology). It is my belief that the methods and software presented in this thesis will be useful for analysing genome-wide diversity, interaction and coevolution with broad applications to many viral infectious diseases.

8.3 HIV-1 protein coevolution

Study of the evolution theory, coevolution is essential for exploring the relationships between species in the complex ecological systems [26]. Recently, sequence-based methods have been established to use the phylogenetic information and to disentangle indirect relationships, leading to an improved prediction capacity [26]. Previous HIV-1 coevolution studies revealed that functional communications of coevolving residues in the GP120-GP41 complex [27] and in the Matrix-GP41 complex [28], both of which help HIV to overcome major structural alterations during viral entry. As reviewed in Chapter 5, more than 27 sequence-based methods have been proposed in

the last decade, making the coevolution research one of the advanced areas in computational biology.

To the best of our knowledge, no computational method has been proven efficient to model genome-wide coevolution. While many methods have been applied to a single protein, I realized that a standard method for investigating the coevolution between different HIV-1 proteins is still lacking. For this reason, I focused on developing a software system, which provided robust predictions of long-range coevolution within and between HIV proteins to the extent of coevolution in the HIV full-length genome. To meet this objective, I designed an ensemble coevolution system (Chapter 5). As my pilot study of HIV-1 genome-wide coevolution, Chapter 6 investigates the HIV-1 Gag-protease coevolution.

Ideas of my HIV coevolution projects

Chapter 5 introduces a new coevolution system that integrates different sequence-based methods to predict amino acid coevolution. While the original goal was to create a method for modeling the genome-wide HIV coevolution, it was not feasible due to several reasons. Firstly, no method has proven efficient in predicting genome-wide residue coevolution. Secondly, no *in vitro* experiment has been developed to identify all the possible amino acid associations in the whole-length genome. Thirdly, biological mechanisms of genome-wide HIV coevolution have not been fully understood.

At the beginning, I tried the structure-based methods, but the prediction results were not interpretable. I also applied many sequence-based methods, but prediction results were largely distinct, causing a problem for data interpretation. Thereafter, I began to investigate the HIV-1 Gag-protease coevolution because many experimental and clinical studies had provided data for the model validation. Designing a new method was still a challenge even though the goal was reset to investigate the HIV-1 Gag-protease coevolution – one of the key coevolution events in the HIV-1 genome. Firstly, even after many hard-working days and nights, I failed to design a single prediction method that could outperform the 30 published methods. Secondly, discordance predictions were obtained from different methods with comparable prediction accuracy, making the data interpretation extremely hard.

Acting on the contention that the performance of sequence-based methods varies, I came up with the idea of assembling different prediction models into one system, which was called “ensemble coevolution system” (Chapter 5). In the meantime, I studied the known algorithms in the research field of ensemble learning [29, 30]. Algorithms such as AdaBoost were however not feasible for assembling the sequence-based methods, because sequence-based methods do not provide negative predictions so that true negatives cannot be evaluated and used in AdaBoost. For this reason, I used the incremental learning and the majority voting to integrate multiple sequence-based methods, resulting in the heuristic algorithm that I proposed in Chapter 5. Using the combination of individual sequence-based methods, combinations of methods could outperform any of the individual methods so that the prediction capacity was improved. Thereafter, I applied the ensemble coevolution system to model the HIV-1 Gag-protease coevolution networks (Chapter 6).

Strength and weakness of my coevolution studies

In Chapter 5 and 6, I proposed a software platform ECS to investigate the HIV-1 Gag-protease coevolution networks, providing valuable insight on HIV genome-wide coevolution. Ensemble coevolution system is the first coevolution system proposed to investigate the residue coevolution given the 27 sequence-based methods. Regarding the weakness of my study, I have only evaluated method performance in HIV-1 proteins. It is possible that different combinations of methods can be more efficient in other protein families. Since most sequence-based coevolution methods have been developed to predict coevolving residues in many protein families [31], as a future perspective, we still need to evaluate the performance of sequence-based methods using large-scale protein family datasets, for instance, using the benchmark datasets provided in the Critical Assessment of protein Structure Prediction (CASP, http://predictioncenter.org/).

Recent studies have demonstrated that an accurate prediction of protein contact maps can improve the prediction of protein 3D structures [32]. In future, we will apply the ensemble coevolution system to predict large protein structures. Since new sequence-based methods are continually being reported, it is also important to integrate new methods in our system for the improvement of prediction capacity. In addition, how to efficiently integrate multiple resources from in vitro experiments and prediction models remains a challenge in the research field of coevolution.

8.4 Probabilistic graphical models

Probabilistic graphical models (e.g. Bayesian networks) have been applied in many fields for the knowledge discovery and inference [33]. Probabilistic graphical models provide powerful tools for biologists to infer cellular networks [34] and to study the possible evolutionary pathways [35]. During my Phd study, my work focused on three types of probabilistic graphical models: Bayesian networks, ancestral polytree models and multi-polytree graphical models.

In Chapter 9, I have proposed novel graphical models called ancestral polytree models, which belong to the ancestral graphical models. The strength of these models relies on the fact that polytree models allow for fast learning and inference when large-scale variables are under investigation. Bayesian networks provide a higher order of probabilistic dependencies in modeling the possible relationships between observed variables. However, modeling large-scale networks demands a high computation, which limits a wide application of Bayesian networks in genome-wide analysis. In Chapter 9, I proposed ancestral polytree models and characterized the properties of ancestral polytrees regarding the Markov properties, the probabilistic factorization and the learning procedure. Using the synthesized and biological datasets, I showed that the promising applications of ancestral polytree models. Further investigations are still needed to solve the maximum likelihood inference in ancestral polytrees.

As extensions of single polytree models described in Chapter 9, I worked on a new class of graphical models, called multi-polytree graphical models (manuscript in preparation). The multi-polytree models can construct a set of polytree models to improve maximum likelihood inference. I proposed the Expectation–maximization algorithm to learn the multi-polytree graphical models and solved the difficulty in identifying the number of variable clusters using fuzzy clustering.

I have worked with Prof. Concha Bielza and Prof. Pedro Larrañaga on the projects of the multi-dimensional Bayesian network classification (MBC) [36]. MBC has been proven useful when multiple target variables are required for robust predictions in many real-word applications [36]. In the past few years, a great number of publications have emerged to propose algorithms that provide accurate predictions of multiple class variables. Compared to single class prediction models (e.g. Naïve Bayesian classifiers), MBC has shown superior performance using both synthetic and

biological datasets. Under supervision from Prof. Concha Bielza and Prof. Pedro Larrañaga, I contributed to data collection, algorithm design and performance evaluations. I also contributed to the analysis of computational complexity of the inference and learning in MBC.

8.5 Data visualization

During my Phd study in an immunology and epidemiology lab, I realized the fact that beautiful mathematical equations are less appreciated than pictures. Although the lesson was difficult to take, the driving forces to improve my skills on data visualization have benefited my research. Here, I would like to share my visualization toolbox and my experience on several data visualization software.

Visualization using my Matlab toolbox: I have developed the toolboxes for visualizing genomic diversity. Basic programming skills are needed to create pictures because the visualization purposes are usually different in individual projects. Examples of how to use my visualization toolbox are available in my Matlab space (http://www.mathworks.com/matlabcentral/fileexchange/authors/45984). Moreover, I worked on the visualization of geographical regions in the world map. While Google Map APIs have provided extensive functions, I have developed the geographical mapping tools in Matlab that provide the flexibility of visualizing the country-specific or region-specific information on the world map.

Visualization of protein structure: I used three tools to visualize protein structures. (1) PyMOL (http://www.pymol.org/), designed by Dr. Warren L. DeLano. As my favorite structural visualization software, this software is user friendly and provides intuitive ways to visualize protein structures. Complex algorithms for making structural movies are available in the software. Various protein structure packages can be easily integrated using Python. (2) Chimera is supported by a team from the University of California, San Francisco (http://www.cgl.ucsf.edu/chimera/). This software is comprehensively documented and has an active user community, which is very helpful to solve problems of most structural analysis. This software is very user-friendly and the implemented structural analyses are very easy to follow for beginners. (3) MOE (http://www.chemcomp.com/) is a famous structural tool utilized in the field of drug design. While many options and algorithms have been implemented in MOE, its license is expensive and the details of implemented algorithms are proprietary.

Among the above three tools, I favor Chimera for protein structural analysis, while PyMOL is better for creating publication pictures. I also tried other tools such as Jmol and Visual molecular dynamics (VMD), but they were not used often in my research.

Visualization of protein-protein interaction (PPI) networks: I used three tools to visualize protein-protein interaction networks. (1) Geomi V2.0 is a 3D visualization software (http://sydney.edu.au/engineering/it/~visual/geomi2/), created by Dr. Seokhee Hong. I used this software to visualize HIV-human protein interaction networks (Chapter 3). This software provides many layout algorithms for the optimization of variable positions – a function which is not commonly optimized. This software requires basic programming skills to prepare the input files for visualizing large-scale datasets. (2) Cytoscape (http://www.cytoscape.org/) is user friendly and provides the impressive visualization of protein-protein interactions. Features of Cytoscape have been published in Nature Method [37]. (3) Graphviz (http://www.graphviz.org/) is a tool used to produce 2D graphical networks. This software is handy to standardize graphical outputs, but basic programming skills are needed to prepare the dot input files. Further reading about different PPI visualization software can be found in a recent article [38].

Visualization of genome-wide interaction networks: Circos (http://circos.ca/) is a visualization tool made by Dr. Martin Krzywinski. The concept of using circles to visualize genome-wide interaction networks is nothing new, but Circos is the best genome-wide visualization tool that I found. In my studies, I used Circos to visualize Gag-protease coevolution networks, HIV genome-wide diversity and HIV-1 inter-protein interaction networks. This software provides many simple and efficient visualization strategies for genome-wide analysis.

Visualization of schematic views of workflow and viral life cycle: I used four tools (Photoshop, Illustrator, Fireworks and PowerPoint) for the visualization of workflows and the viral life cycle. PowerPoint is my favorite software and provides simple and quick options for making pictures. Photoshop and Illustrator are professional visualization software, but require more time and training to master skills.

Visualization of phylogenetic trees: Many visualization tools have been developed, but FigTree (http://tree.bio.ed.ac.uk/software/figtree/) has received the most attention.

Many nice features are provided in FigTree, such as the highlight of certain branches and various shapes of phylogenetic trees.

8.6 Future perspectives

8.6.1 Genome-wide interactions between HIV and human proteins

More than 1000 human proteins have been found to interact with HIV proteins during the viral life cycle [39]. Based on the literature results, possible HIV-human protein interactions have been classified by the NCBI HIV-human protein interaction database (http://www.ncbi.nlm.nih.gov/projects/RefSeq/HIVInteractions/). According to the information extracted from the above database, human proteins can upregulate, downregulate, inhibit, disrupt, bind, inactivate, stimulate, enhance, activate, polarize, methylate, ubiquitinate, myristoylate, phosphorylate, dephosphorylate, deglycosylate, depolymerize, stabilize, cleave, re-localize, co-localize, export or import HIV proteins during viral infection and replication. The global landscape of HIV-human interaction system has thus been reconstructed for more than 1500 human proteins, which directly (e.g. bind) or indirectly (e.g. upregulate) interact with HIV proteins [40]. For instance, the human protein PAF1 [41] and APOBEC3G [42] can prevent HIV infection through HIV-human protein interactions during the viral life cycle. Understanding HIV-human protein interactions can be useful in the development of HIV drugs and vaccines.

The NCBI HIV-human protein database has been criticized by the low quality of literature data, which contains a certain proportion of false-positive interactions due to different date ranges, a diverse array of experimental procedures, inconsistent and redundant annotations of interactions [43]. Different HIV-human protein interactions have also been reported in different subtypes, while this difference is hardly distinguishable in NCBI database. Several independent studies have been proposed to identify HIV-human protein interactions based on the siRNA gene knockdown screens [44, 45]. Despite this, a low coverage between different studies has been reported, largely due to differences in experimental procedures (e.g. cell-type, choice of time points analyzed and choice of filtering thresholds) [43]. In addition, siRNA-based studies could not infer the detailed information about residue positions that are responsible for HIV-human protein interactions. Structures of HIV-human protein interaction are mostly lacking, nor the protein interaction positions.

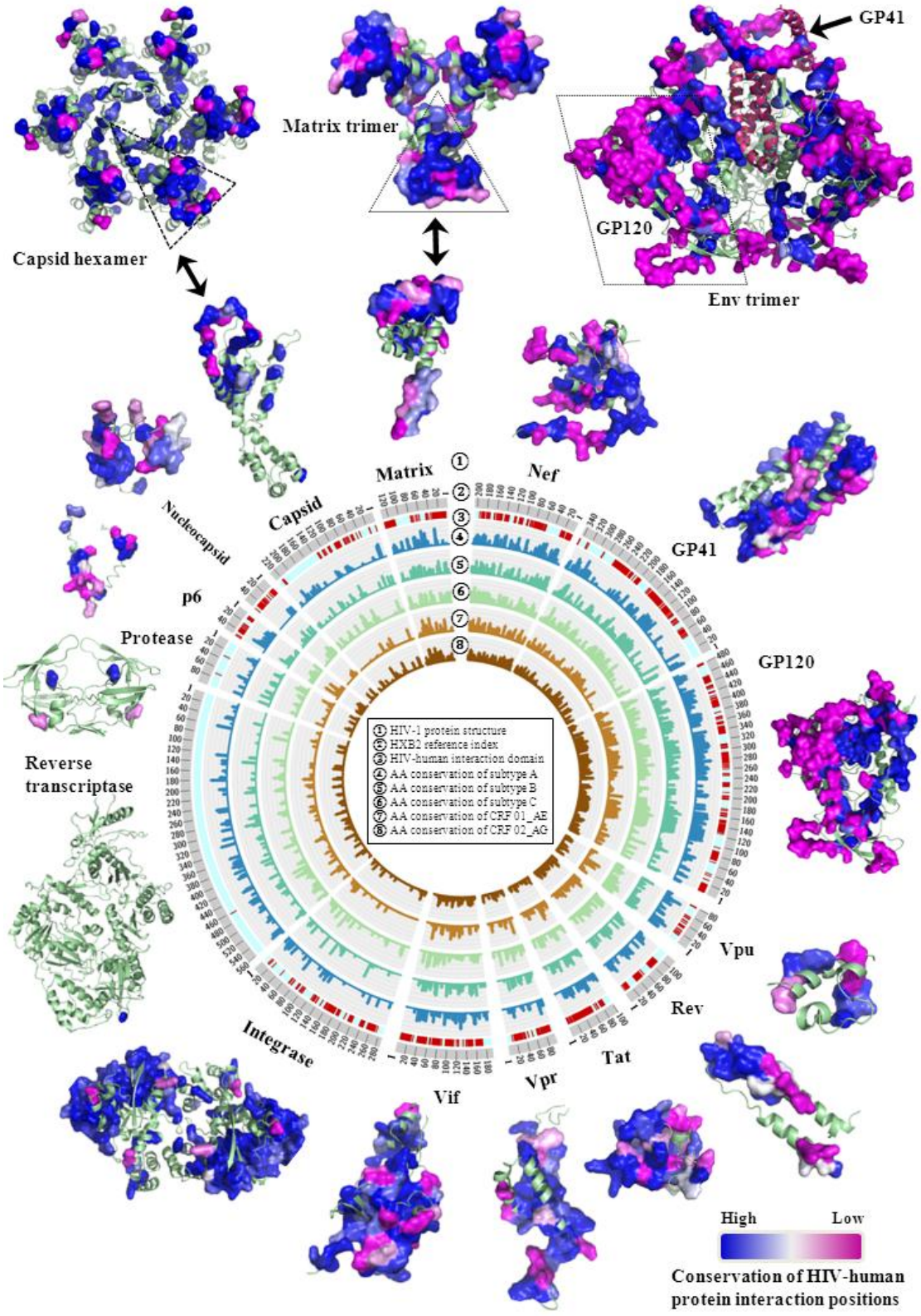


Figure 8.1: HIV genome-wide conservation and HIV-human protein interaction domains. Layer 1: HIV-1 protein structures with the surface representation of HIV-human protein interaction domains. Layer 2: HBX2 reference index. Layer 3: mapping of the known HIV-human protein interaction positions. Layer 4-8: distributions of amino acid conservation index at the full-length genome across 5 major HIV-1 subtypes (A, B, C) and CRFs (01_AE,

02_AG). The list of PDB data is available in Chapter 3. Visualization software: PyMOL V1.5 (http://www.pymol.org/), Circos V0.64 (http://circos.ca/).

The NCBI HIV-human protein database has inspired me to understand the HIV-human protein interactions. However, many redundant and confusing annotations have been observed in this database (approximately 5-10% based on my personal estimation). This database has indeed provided valuable information but scientifically, the data quality still needs improvement. While the NCBI database has been criticized by many false-positive protein interactions [43], previous siRNA-based studies only reported possible protein-protein interactions but not the interaction domains.

I performed literature review to collect information about the HIV-human protein interactions and their protein interaction positions reported in experiments. Based on HIV publications in the last three decades, I conducted the literature review over 5000 publications and had selected about 700 publications showing the experimental data of direct interaction positions between HIV and human proteins. This process took me a few months and enriched my knowledge about HIV and human proteins, which in turn enlightened me to outline my future project which aimed at identifying the human proteins that physically interact with HIV proteins during the viral life cycle. As a preliminary visualization result in **Figure 8.1**, the known HIV-human protein interaction positions and their amino acid conservation are mapped in the 5 major HIV-1 subtype and CRF genomes using 94560 sequences. Information of HIV-human protein interaction positions and amino acid sequence datasets has also been included in **Appendix 1 and 2**. Further work still need to investigate how human genomes coevolve with HIV genome, leading to one of the most complex virus-host interaction networks being identified.

8.6.2 Comparison of HIV, HBV and HCV genomic diversity

Inspired by my HIV projects, I will briefly highlight two side projects about HBV and HCV that I will cooperate with colleagues in my research lab.

HCV is a RNA virus and belongs to the genus Hepacivirus in the Flaviviridae family. A HCV particle contains a single-stranded RNA genome that encodes 10 proteins in one open reading frame [46]. The HCV RNA genome has 9000 to 9100 nucleotides

depending on the HCV genotypes. Moreover, ten HCV proteins are classified as structural and non-structural proteins. HCV structural proteins include one core (Core) and two envelope proteins (E1, E2). HCV non-structural proteins include NS2, NS3, NS4A, NS4B, NS5A and NS5B. Notably, the NS3 protease and the NS5B polymerase are two important viral enzymes. Besides the structural and non-structural proteins, P7 is a membrane-associated oligomeric protein, whose ion channel activity is crucial for the assembly of HCV particles [46].

HBV is a DNA non-retroviral virus in the hepadnavirus family and has a circular genome of partially double-stranded DNA. HBV particles contain one full-length strand (3020 – 3300 nucleotides) and one short-length strand (1700 – 2800 nucleotides)[47]. Four genes (P, S, C, X) in the HBV genome encode for viral proteins which play multiple roles during the HBV life cycle. (1) The HBV polymerase gene P encodes for the DNA polymerase which synthesizes DNA molecules by assembling nucleotides in the cytoplasm. (2) The S gene encodes for the surface antigen (HBsAg), which contains three sections (pre-S1, pre-S2, S). (3) The HBV core gene C encodes for the Core and Pre-C protein which construct the core of HBV particles. (4) The X gene encodes for a protein whose functions have not been fully understood [48].

I used the same method in Chapter 3 to quantify the nucleotide genomic diversity of HIV, HBV and HCV genomes. As my preliminary result visualized in **Figure 8.2**, two major observations were identified. (1) When comparing the intra-subtype diversity, HBV has the lowest genomic diversity compared to HCV and HIV. The intra-subtype diversity of HCV subtype 1b is comparable to HIV-1 subtypes, while HCV subtype 1a has a lower genomic diversity than HIV-1 subtypes. (2) When comparing the inter-subtype diversity, HCV has the highest inter-subtype genomic diversity compared to HBV and HIV. Distribution of HBV inter-subtype genomic diversity has two major peaks; the one with higher diversity is comparable to HIV-1 inter-subtype genomic diversity. HCV inter-subtype genomic diversity has a similar distribution as the genomic diversity between HIV-1 group M and O/P. The genomic diversity between HIV-1 and HIV-2 is the highest among all analysed subtypes, groups and types in HBV, HCV and HIV.

Overall, our comparative analysis suggests that the genomic diversity of three viruses follows the order as: HIV > HCV > HBV.

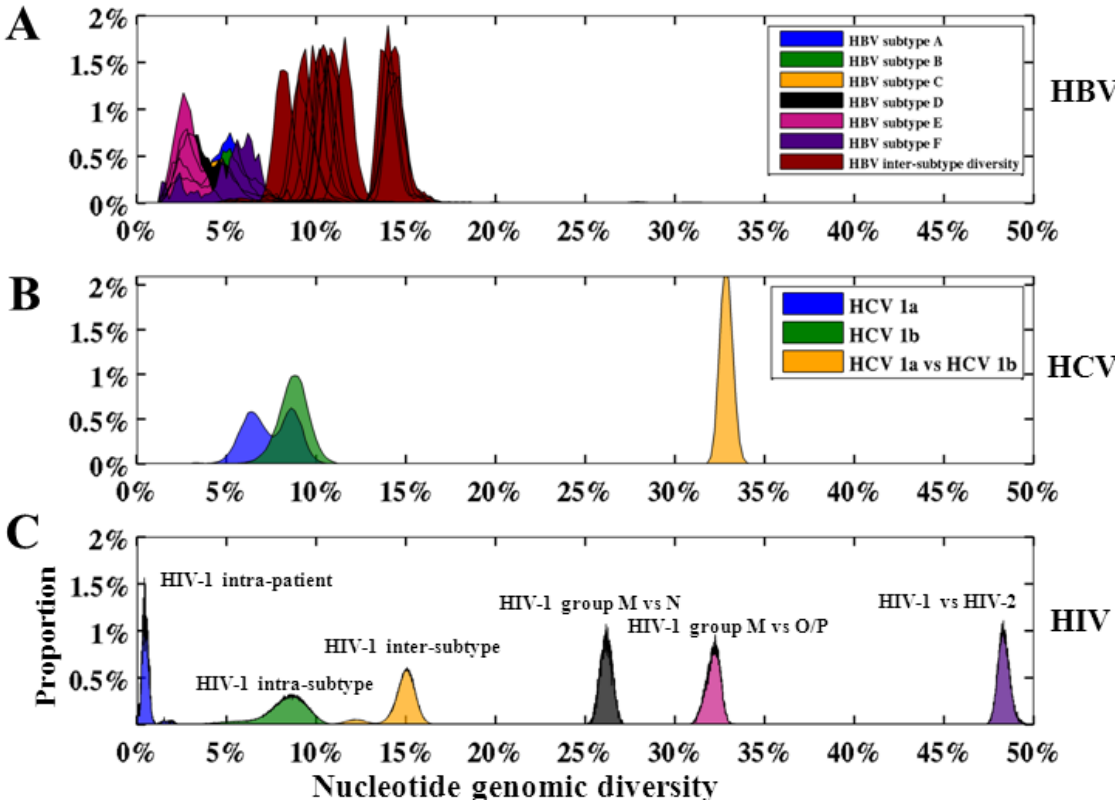


Figure 8.2: Distribution of HBV (A), HCV (B) and HIV (C) genomic diversity. The x- and y-axes demonstrate nucleotide genomic diversity and proportion, respectively. Different viral subtypes have been colored indicated by the figure legend.

8.7 Author's words in the end

When I look back to my Phd study wondering around in the research area, I can not remember how many times I got excited with small results and how many times I got upset when my analyses failed. However, "Failures are finger posts on the road to achievement." – C. S. lewis.

In fact, most of my projects failed during my Phd study. For instance, I failed to design a single sequence-based method to outperform all published methods. I was unable to design a software platform that integrates sequence and protein structural analysis. I failed to construct the structure of HIV integrase tetramer. I could not design any peptide or molecular inhibitor after many theoretical tries. I was unable to model the coevolution between HIV and human genomes. Although many of my projects have failed during my Phd, some might be achievable in the future. Many things that I wanted to do during my Phd but I could not do. For you, who can see any

value of this thesis, I wish you could stand on my shoulder to become a giant: "If I have seen further it is by standing on the shoulders of giants" – Isaac Newton.

My Phd thesis cannot be successfully presented without substantial advices from my Phd promoter, co-promoters, colleagues and jury members. It is therefore my sincere acknowledgement for their contributions.

8.8 References

1. Clavel F, Hance AJ. HIV drug resistance. *New England Journal of Medicine* 2004,**350**:1023-1035.
2. Tang MW, Shafer RW. HIV-1 antiretroviral resistance: scientific principles and clinical applications. *Drugs* 2012,**72**:e1-25.
3. Nijhuis M, van Maarseveen NM, Lastere S, Schipper P, Coakley E, Glass B, *et al.* A novel substrate-based HIV-1 protease inhibitor drug resistance mechanism. *PLoS Med* 2007,**4**:e36.
4. Fun A, Wensing AM, Verheyen J, Nijhuis M. Human Immunodeficiency Virus Gag and protease: partners in resistance. *Retrovirology* 2012,**9**:63.
5. van Sighem A, Gras L, Reiss P, Brinkman K, de Wolf F. Life expectancy of recently diagnosed asymptomatic HIV-infected patients approaches that of uninfected individuals. *Aids* 2010,**24**:1527-1535.
6. Li G, Verheyen J, Rhee SY, Voet A, Vandamme AM, Theys K. Functional conservation of HIV-1 gag: implications for rational drug design. *Retrovirology* 2013,**10**:126.
7. Adamson CS, Sakalian M, Salzwedel K, Freed EO. Polymorphisms in Gag spacer peptide 1 confer varying levels of resistance to the HIV- 1 maturation inhibitor bevirimat. *Retrovirology* 2010,**7**:36.
8. Larrouy L, Vivot A, Charpentier C, Benard A, Visseaux B, Damond F, *et al.* Impact of gag genetic determinants on virological outcome to boosted lopinavir-containing regimen in HIV-2-infected patients. *AIDS* 2013,**27**:69-80.
9. Hemelaar J. The origin and diversity of the HIV-1 pandemic. *Trends Mol Med* 2012,**18**:182-192.
10. Rolland M, Tovanabutra S, deCamp AC, Frahm N, Gilbert PB, Sanders-Buell E, *et al.* Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. *Nat Med* 2011,**17**:366-371.
11. Haynes BF, Gilbert PB, McElrath MJ, Zolla-Pazner S, Tomaras GD, Alam SM, *et al.* Immune-correlates analysis of an HIV-1 vaccine efficacy trial. *N Engl J Med* 2012,**366**:1275-1286.
12. Rolland M, Edlefsen PT, Larsen BB, Tovanabutra S, Sanders-Buell E, Hertz T, *et al.* Increased HIV-1 vaccine efficacy against viruses with genetic signatures in Env V2. *Nature* 2012,**490**:417-420.
13. Herbeck JT, Rolland M, Liu Y, McLaughlin S, McNevin J, Zhao H, *et al.* Demographic processes affect HIV-1 evolution in primary infection before the onset of selective processes. *J Virol* 2011,**85**:7523-7534.
14. Rousseau CM, Birditt BA, McKay AR, Stoddard JN, Lee TC, McLaughlin S, *et al.* Large-scale amplification, cloning and sequencing of near full-length HIV-1 subtype C genomes. *J Virol Methods* 2006,**136**:118-125.
15. Tee KK, Li XJ, Nohtomi K, Ng KP, Kamarulzaman A, Takebe Y. Identification of a novel circulating recombinant form (CRF33_01B) disseminating widely among various risk populations in Kuala Lumpur, Malaysia. *J Acquir Immune Defic Syndr* 2006,**43**:523-529.
16. Arien KK, Vanham G, Arts EJ. Is HIV-1 evolving to a less virulent form in humans? *Nat Rev Microbiol* 2007,**5**:141-151.
17. Guyader M, Emerman M, Sonigo P, Clavel F, Montagnier L, Alizon M. Genome organization and transactivation of the human immunodeficiency virus type 2. *Nature* 1987,**326**:662-669.
18. Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C, Detours V. Evolutionary and immunological implications of contemporary HIV-1 variation. *Br Med Bull* 2001,**58**:19-42.

19. Wang YE, Li B, Carlson JM, Streeck H, Gladden AD, Goodman R, *et al.* Protective HLA class I alleles that restrict acute-phase CD8+ T-cell responses are associated with viral escape mutations located in highly conserved regions of human immunodeficiency virus type 1. *J Virol* 2009,**83**:1845-1855.
20. Brown BK, Darden JM, Tovanabutra S, Oblander T, Frost J, Sanders-Buell E, *et al.* Biologic and genetic characterization of a panel of 60 human immunodeficiency virus type 1 isolates, representing clades A, B, C, D, CRF01_AE, and CRF02_AG, for the development and assessment of candidate vaccines. *J Virol* 2005,**79**:6089-6101.
21. Fernandez-Garcia A, Cuevas MT, Munoz-Nieto M, Ocampo A, Pinilla M, Garcia V, *et al.* Development of a panel of well-characterized human immunodeficiency virus type 1 isolates from newly diagnosed patients including acute and recent infections. *AIDS Res Hum Retroviruses* 2009,**25**:93-102.
22. Kousiappa I, Van De Vijver DA, Kostrikis LG. Near full-length genetic analysis of HIV sequences derived from Cyprus: evidence of a highly polyphyletic and evolving infection. *AIDS Res Hum Retroviruses* 2009,**25**:727-740.
23. Sanabani SS, Pessoa R, Soares de Oliveira AC, Martinez VP, Giret MT, de Menezes Succi RC, *et al.* Variability of HIV-1 genomes among children and adolescents from Sao Paulo, Brazil. *PLoS One* 2013,**8**:e62552.
24. Fernandez-Garcia A, Revilla A, Vazquez-de Parga E, Vinogradova A, Rakhmanova A, Karamov E, *et al.* The analysis of near full-length genome sequences of HIV type 1 subtype A viruses from Russia supports the monophyly of major intrasubtype clusters. *AIDS Res Hum Retroviruses* 2012,**28**:1340-1343.
25. Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, *et al.* Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog* 2012,**8**:e1002529.
26. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. *Nat Rev Genet* 2013,**14**:249-261.
27. Travers SA, Tully DC, McCormack GP, Fares MA. A study of the coevolutionary patterns operating within the env gene of the HIV-1 group M subtypes. *Mol Biol Evol* 2007,**24**:2787-2801.
28. Beaumont E, Vendrame D, Verrier B, Roch E, Biron F, Barin F, *et al.* Matrix and envelope coevolution revealed in a patient monitored since primary infection with human immunodeficiency virus type 1. *J Virol* 2009,**83**:9875-9889.
29. Dietterich TG. Ensemble methods in machine learning. In: *Multiple classifier systems*: Springer; 2000. pp. 1-15.
30. Polikar R. Ensemble learning. In: *Ensemble Machine Learning*: Springer; 2012. pp. 1-34.
31. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A* 2011,**108**:E1293-1301.
32. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* 2014,**3**:e02030.
33. Koller D, Friedman N. *Probabilistic graphical models: principles and techniques*: MIT press; 2009.
34. Friedman N. Inferring cellular networks using probabilistic graphical models. *Science* 2004,**303**:799-805.
35. Deforche K, Camacho R, Grossman Z, Silander T, Soares MA, Moreau Y, *et al.* Bayesian network analysis of resistance pathways against HIV-1 protease inhibitors. *Infect Genet Evol* 2007,**7**:382-390.
36. Bielza C, Li G, Larranaga P. Multi-dimensional classification with Bayesian networks. *International Journal of Approximate Reasoning* 2011,**52**:705-727.
37. Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, *et al.* A travel guide to Cytoscape plugins. *Nat Methods* 2012,**9**:1069-1076.
38. Agapito G, Guzzi PH, Cannataro M. Visualization of protein interaction networks: problems and solutions. *BMC Bioinformatics* 2013,**14 Suppl 1**:S1.
39. Jager S, Cimermancic P, Gulbahce N, Johnson JR, McGovern KE, Clarke SC, *et al.* Global landscape of HIV-human protein complexes. *Nature* 2012,**481**:365-370.
40. Fu W, Sanders-Beer BE, Katz KS, Maglott DR, Pruitt KD, Ptak RG. Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Res* 2009,**37**:D417-422.

41. Liu L, Oliveira NM, Cheney KM, Pade C, Dreja H, Bergin AM, *et al.* A whole genome screen for HIV restriction factors. *Retrovirology* 2011,**8**:94.
42. Marin M, Rose KM, Kozak SL, Kabat D. HIV-1 Vif protein binds the editing enzyme APOBEC3G and induces its degradation. *Nat Med* 2003,**9**:1398-1403.
43. MacPherson JI, Dickerson JE, Pinney JW, Robertson DL. Patterns of HIV-1 protein interaction identify perturbed host-cellular subsystems. *PLoS Comput Biol* 2010,**6**:e1000863.
44. Dyer MD, Murali TM, Sobral BW. The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathog* 2008,**4**:e32.
45. Konig R, Zhou Y, Elleder D, Diamond TL, Bonamy GM, Irelan JT, *et al.* Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell* 2008,**135**:49-60.
46. Bartenschlager R, Penin F, Lohmann V, Andre P. Assembly of infectious hepatitis C virus particles. *Trends Microbiol* 2011,**19**:95-103.
47. Kay A, Zoulim F. Hepatitis B virus genetic variability and evolution. *Virus Res* 2007,**127**:164-176.
48. Seeger C, Mason WS. Hepatitis B virus biology. *Microbiol Mol Biol Rev* 2000,**64**:51-68.

Chapter 9

Appendix

“Success is getting what you want, happiness is wanting what you get.”

— Benjamin Franklin

Appendix 9.1: Natural variations in the HIV-1 genome

9.1 Summary of natural variations in the HIV-1 genome

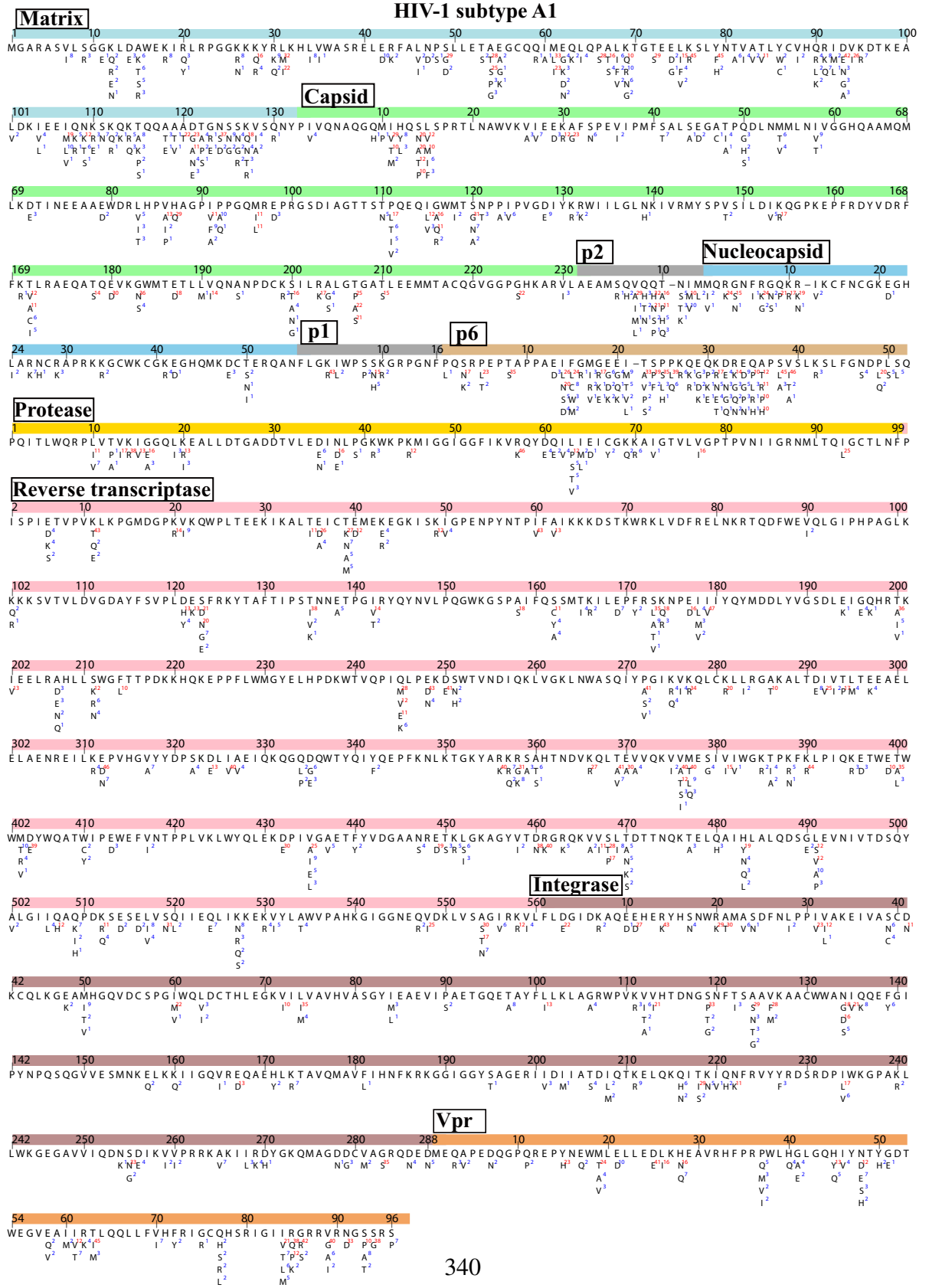
In April 2014, I extracted 213428 HIV-1 sequences from the HIV Los Alamos database (www.hiv.lanl.gov/) to investigate the prevalence of natural variations in large-scale HIV-1 populations. One sequence per patient was extracted and these sequences contained at least one protein region. The quality criteria for removing misclassified sequences or sequences with hypermutations, stop codons, ambiguous nucleotides were described in Chapter 2. I also removed sequences conferred partial or full resistance to any of the protease inhibitors, RT inhibitors and integrase inhibitors using HIVdb V6.0 (http://sierra2.stanford.edu/sierra/servlet/JSierra). I further selected those subtype datasets that contained at least 50 sequences for each HIV-1 protein, resulting in 5 HIV-1 datasets (A1, B, C, D, CRF01_AE, CRF02_AG) (**Table 9.1**) with 94560 sequences in total.

Table 9.1: Summary of HIV-1 sequence datasets

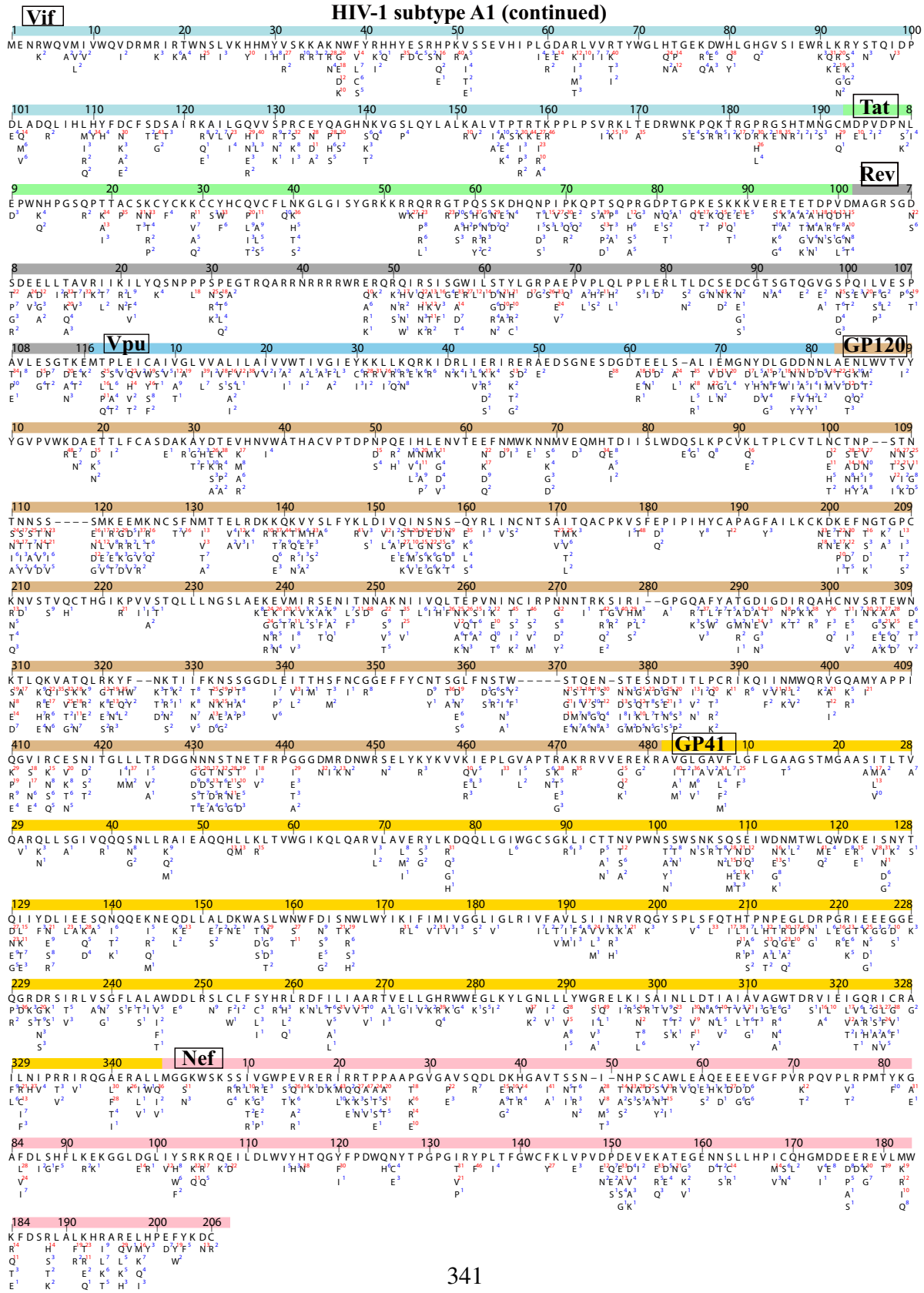
	MA	CA	p2	NC	p1	p6	PR	RT	IN	Vif	Vpr	Tat	Rev	Vpu	GP120	GP41	Nef	Total
A1	784	703	874	381	833	1046	4946	164	233	183	162	172	182	338	315	286	207	7126
B	4725	4517	4068	3748	5494	5177	42070	2357	2510	1962	1847	2475	2619	1777	2302	2171	2434	56891
C	2401	2190	2396	1929	2964	2733	11682	753	617	498	557	882	953	708	1539	1480	806	17149
01_AE	1562	1316	1377	1342	1750	1784	5858	801	834	378	372	384	390	428	610	591	332	7851
02_AG	133	371	633	187	659	712	4279	171	231	101	82	79	91	93	149	141	136	5543

The figures in the following pages visualize natural amino acid variations at all positions of HIV-1 proteins. For each subtype, two pictures are available with the mapping of 15 HIV-1 proteins. The first amino acid position of each protein region is labeled with its protein name in a box. Annotated protein regions are shown as colored bars. HXB2 indices of individual proteins are shown on top of the colored bars. For each subtype, a consensus amino acid at each position is shown beneath the colored bar. Natural variations are shown below the consensus amino acids; proportions (%) are colored red if they were more than 5%; blue otherwise.

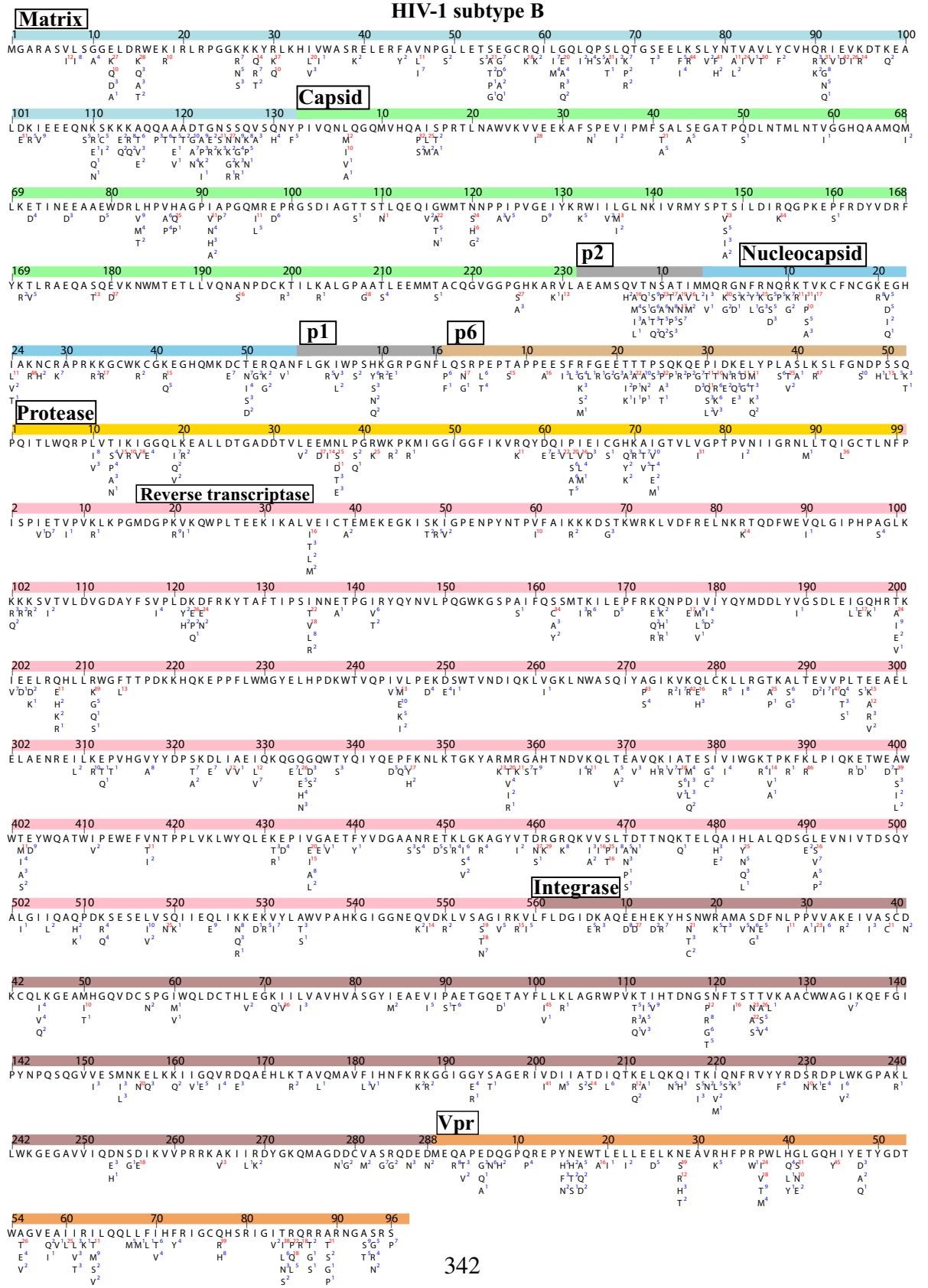
Appendix 9.1: Natural variations in the HIV-1 genome



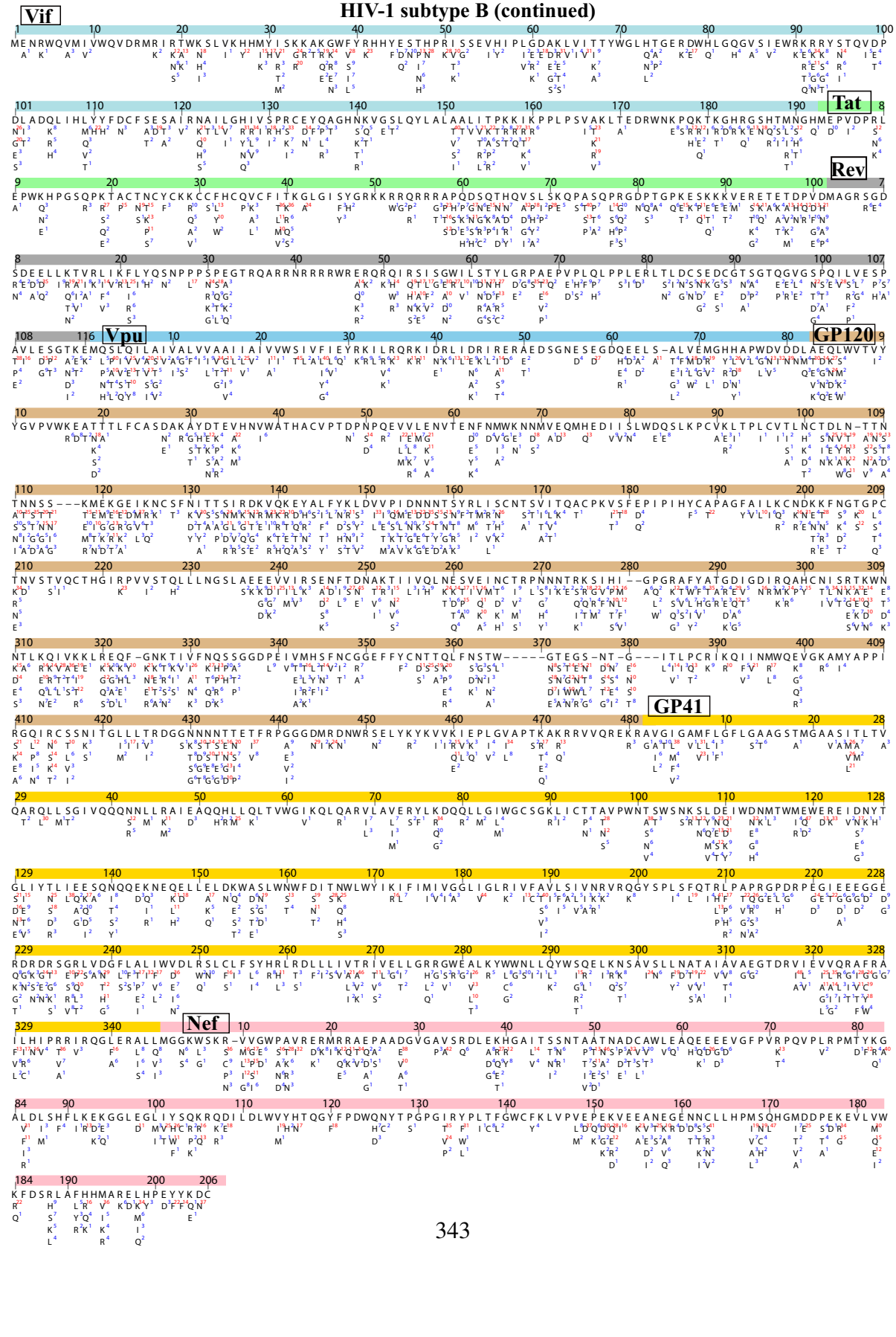
Appendix 9.1: Natural variations in the HIV-1 genome



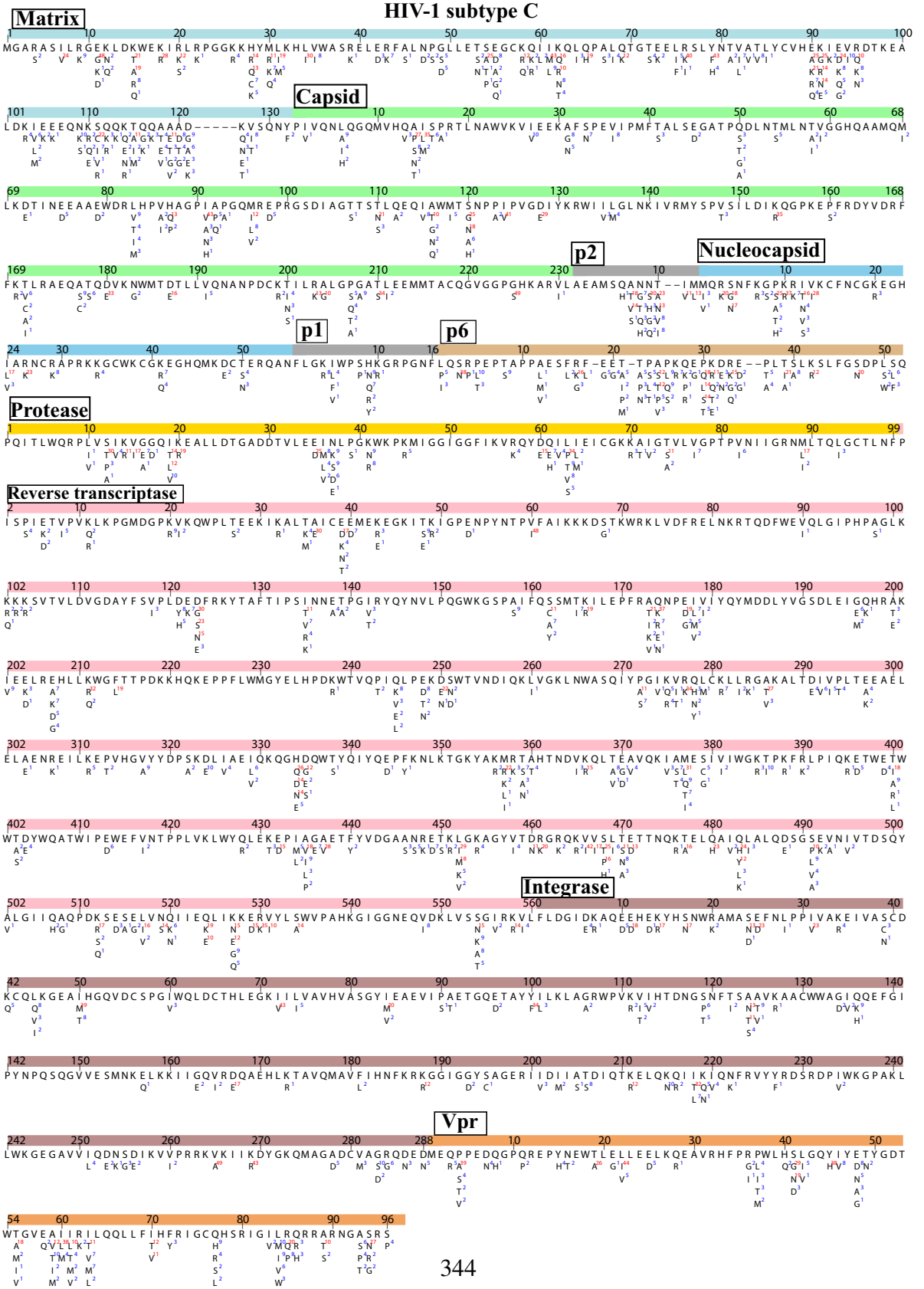
Appendix 9.1: Natural variations in the HIV-1 genome



Appendix 9.1: Natural variations in the HIV-1 genome

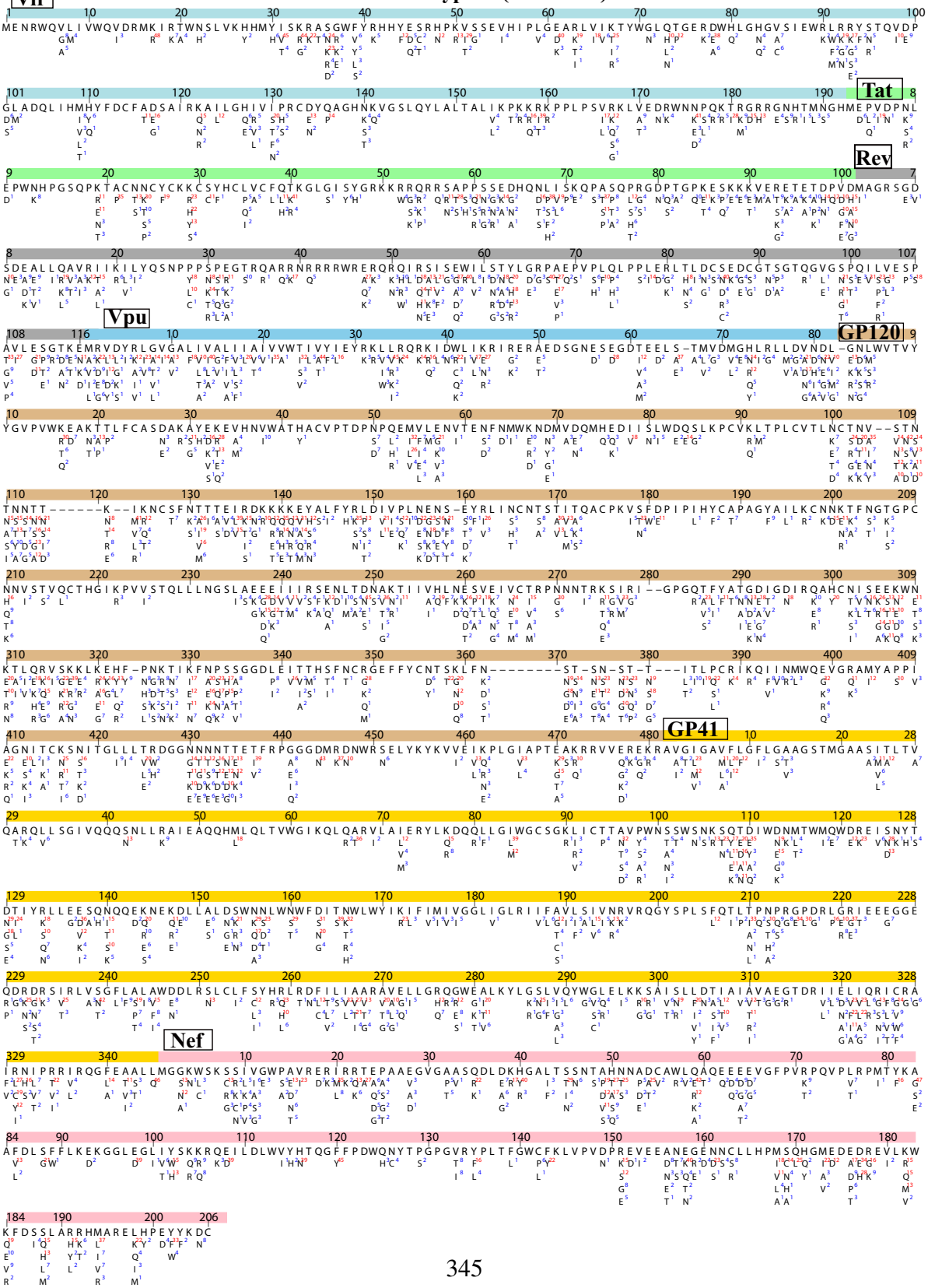


Appendix 9.1: Natural variations in the HIV-1 genome

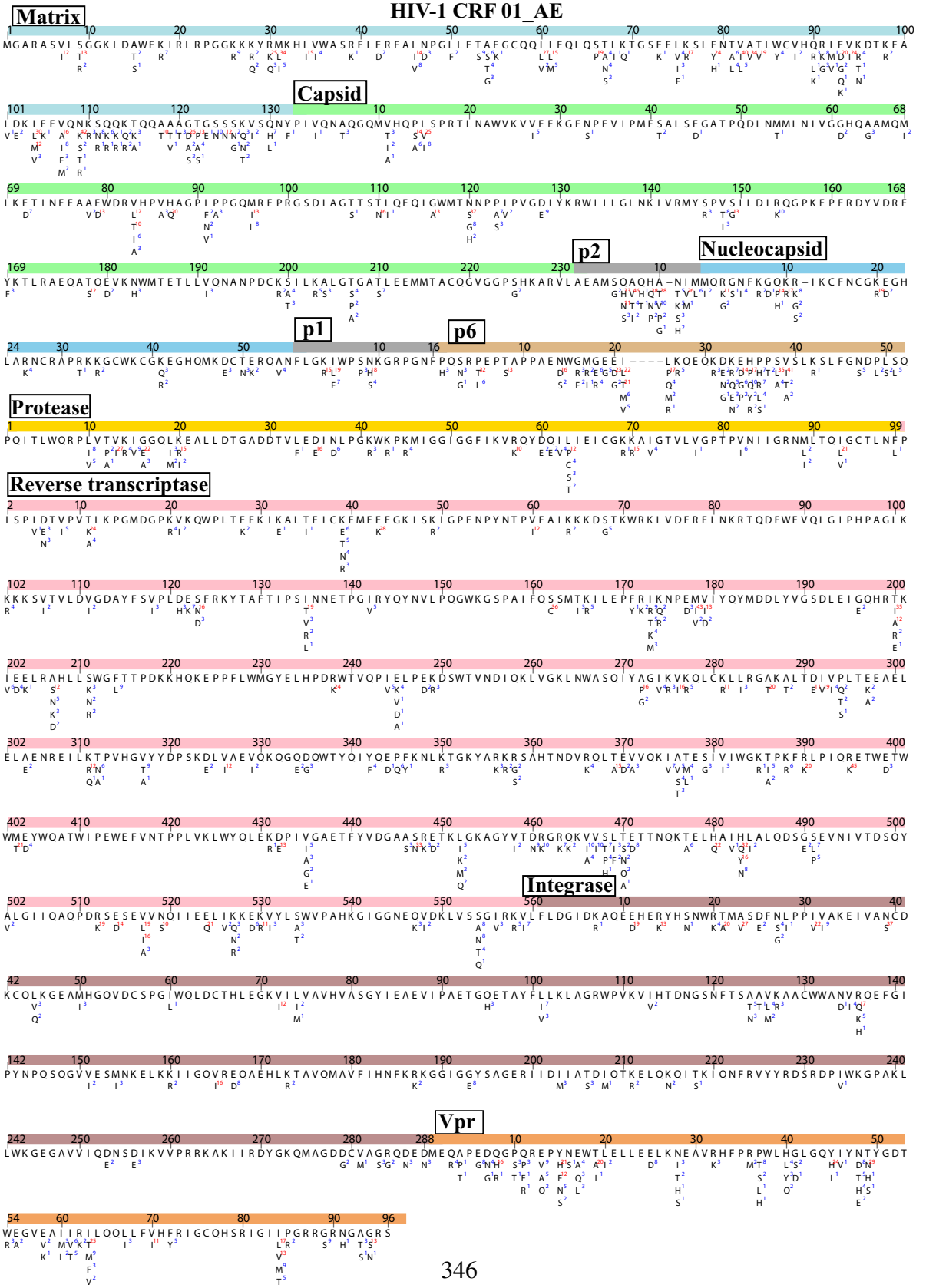


Appendix 9.1: Natural variations in the HIV-1 genome

HIV-1 subtype C (continued)

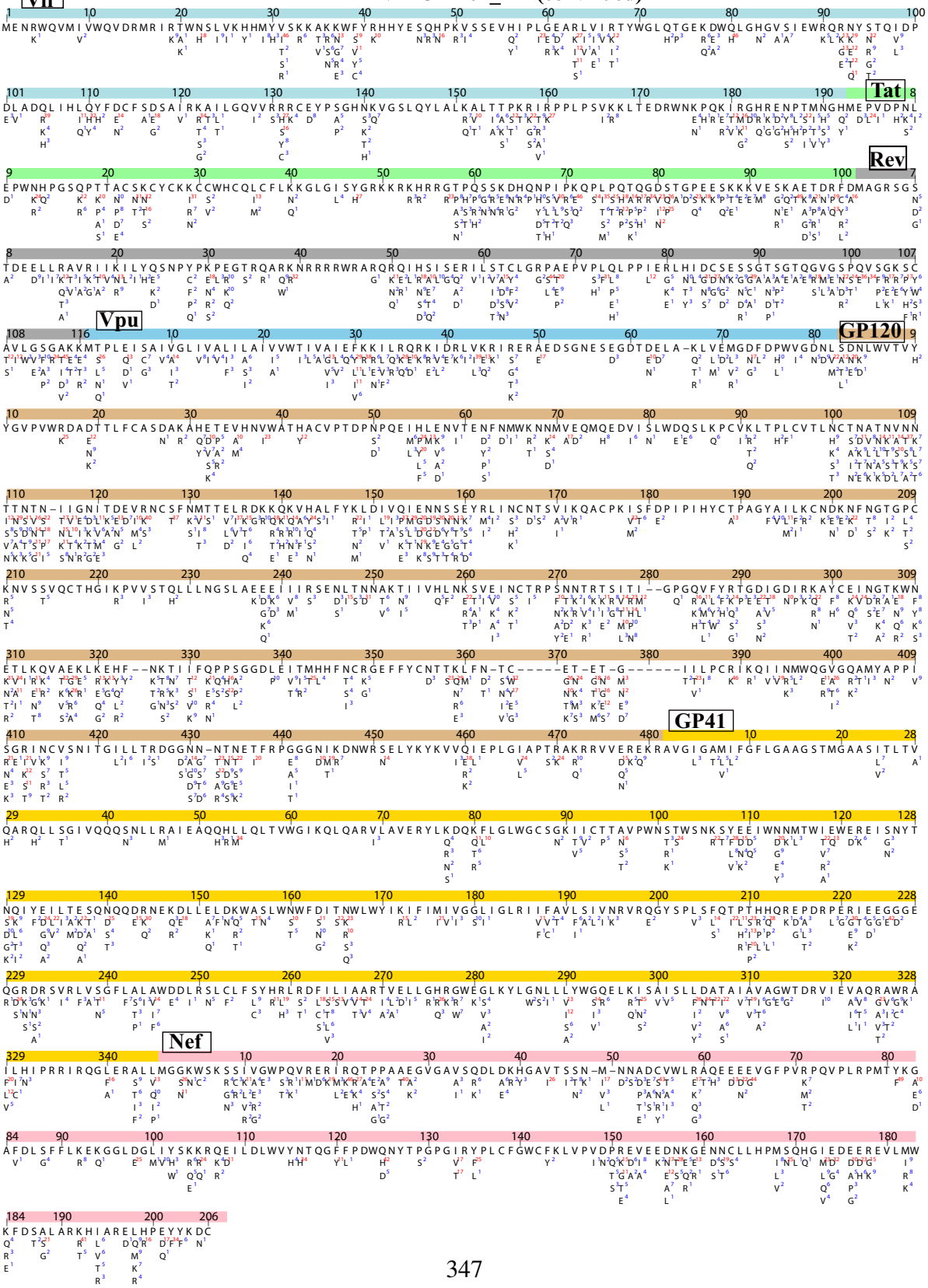


Appendix 9.1: Natural variations in the HIV-1 genome

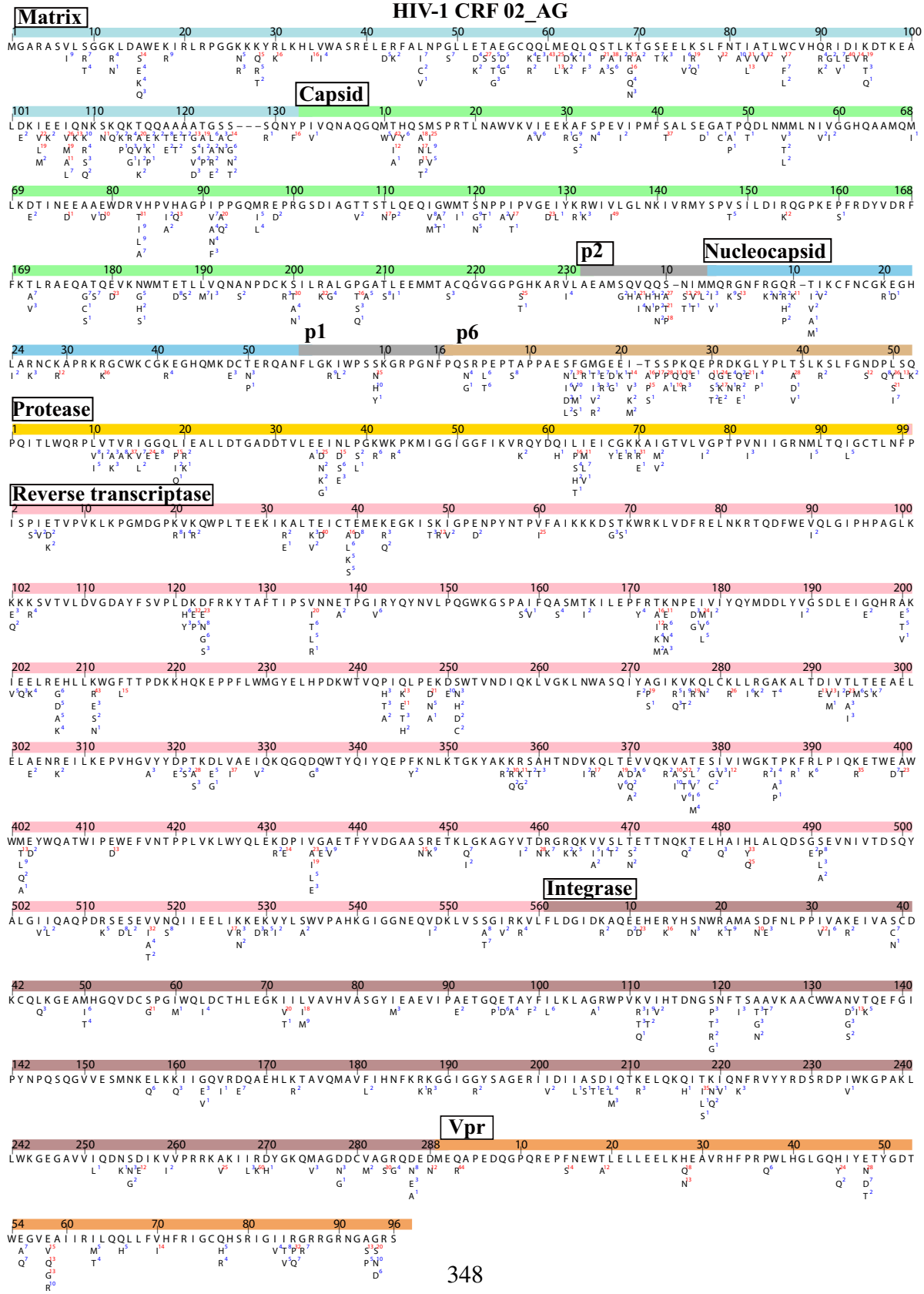


Appendix 9.1: Natural variations in the HIV-1 genome

HIV-1 CRF 01_AE (continued)

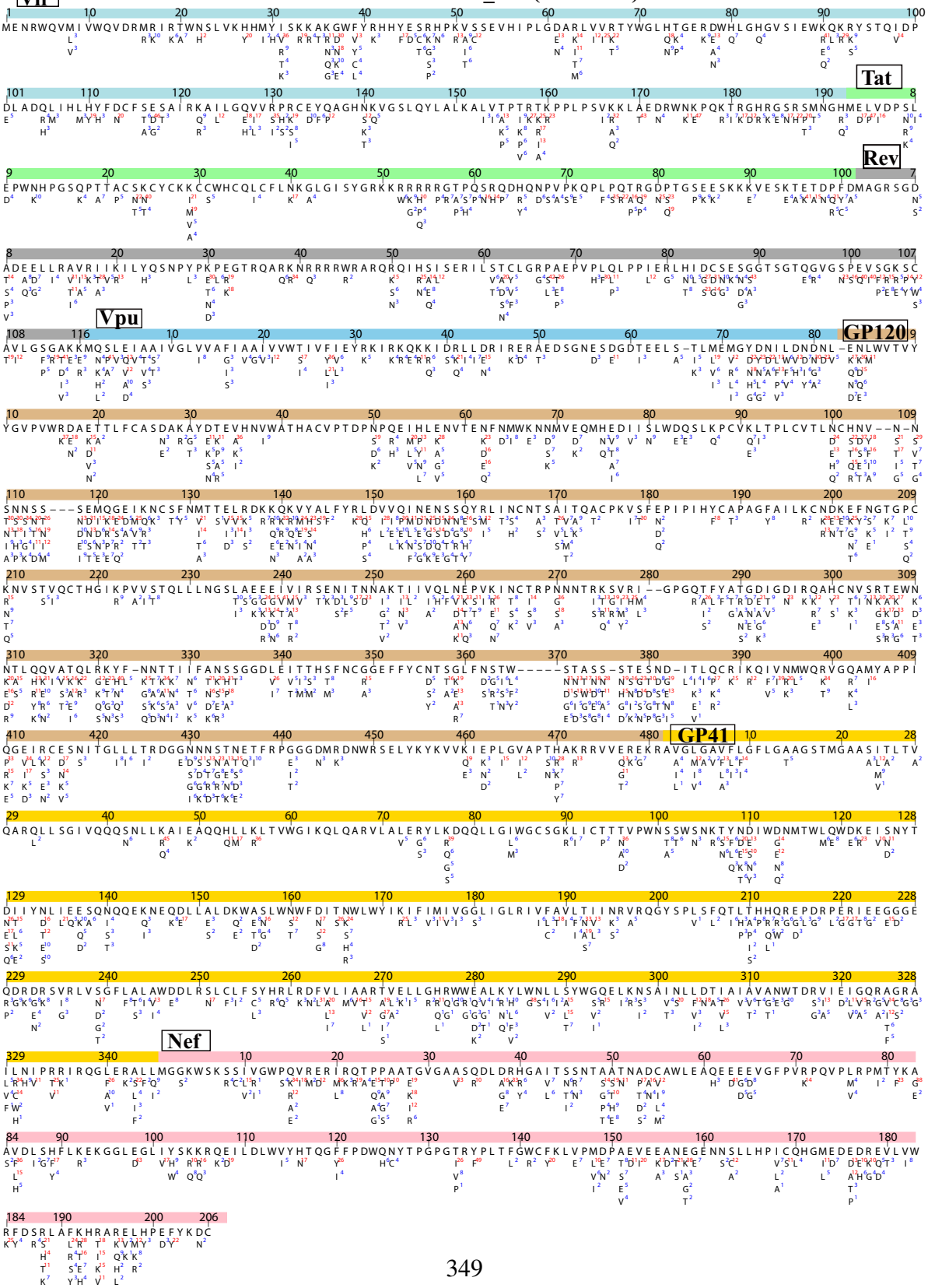


Appendix 9.1: Natural variations in the HIV-1 genome



Appendix 9.1: Natural variations in the HIV-1 genome

HIV-1 CRF 02_AG (continued)



9.2 Summary of HIV-human protein interactions

I performed literature review to collect information on HIV-human protein interactions. Publication selection criteria include: (1) direct HIV-human protein interactions must be reported; (2) interaction positions must be reported either in HIV proteins or in human proteins; (3) results in the latest publication are retained if disagreed results were reported. Subtype-specific information was also collected if it were available. Notably, HIV-human protein interactions may be subtype-specific (e.g. the Vpu-tetherin interaction occurs in group M, but not in group O [1, 2]). The HXB2 reference is used to index the HIV-human interaction positions in HIV proteins.

Table S9.2: Summary of HIV-human protein interactions

HIV proteins	HIV binding positions (1)	Host factors in human	Human protein binding domains	Subtype	Interaction function	Reference
Matrix	8-43[3], W16, W36 [4]	ubiquitous calcium-sensing calmodulin (CaM, CAMI)	F19,G33,L48,E54,V 55,A57,F68,E82,L10 5,T110[5]	HIV-1	Gag intracellular Trafficking in cytoplasm	[3],[4],[5],[6]
	R4-L13,R20-E40[7]	embryonic ectoderm development (EED) protein	Q294-N309[8], 388-403[7]	Subtype B	transcriptional regulation within the nucleus[7]	[7],[8]
	S9-K28	IL-8 chemokine receptors CXCR1(IL-8RA)/CXCR2		HIV-1	endothelial cells proangiogenic activation, monocyte migration	[9],[10],[11]
	K26-K28	elongation factor 1-alpha (EF1a)	1-74	Subtype B	inhibition of translation	[12]
	121-132	Cyclophilin A(CypA)		HIV-1	Enhance CA- CypA interaction	[13]
	15-32	Chromosome maintenance region 1 (CRM1,exportin 1, XPO1)		Subtype B	MA nuclear export signal	[14]
	24-31, 110-114[15]	Importin α1 (karyopherin α2, Rch1/SRP1α/KPNA2)		HIV-1	PIC nuclear import to nucleus	[15]
	Y132	clathrin adaptor complex 2 AP-2, μ2 subunit (AP50)		HIV-1	Gag intracellular Trafficking	[16]
	5-8,13-16	TIP47 (tail-interacting protein of 47 kDa, perilipin 3, PLIN3)		HIV-1	Env packaging into virions	[17],[18]
	K26,K27	HO3 histidyl-tRNA synthetase (HARS2)		HIV-1	HO3 packaging into virions	[19]
	9,67,72,77[20]	mitogen-activated protein kinase (MAPK/ERK-2)	201-255[21]	HIV-1, SIV	Phosphorylation of matrix during early uncoating in cytoplasm, MAPK/ERK-2 incorporation to virions	[20],[22],[21]
	L41,F44,V46,I60,L6 4,L75	PS/PE/PC (phosphatidylserine, phosphatidylethanolamine,phosph atidylcholine)		HIV-1	Gag binding to membrane	[23],[24]
	113-122	Mab 3H7 antibody		HIV-1	PIC integration	[25]
	S111	protease kinase C (PKC)		Subtype B	matrix translocation to membrane	[26],[27]
	25-34,109-115	Heparin(HSPG analog)		HIV-1	prevents p17 binding to chemochine receptor	[11]
	86-115	neutralizing monoclonal antibody (MAb 1575)				[28]
	(HIV-1)S6[29], R22, K26, K27, W36, R76[30],(HIV-2)L21,R22,K27,H33 ,W35,K76,S77[31]	phosphatidylinositol-(4,5)-bisphosphate PI(4,5)P2		HIV-1, HIV-2	targeting Gag to membrane rafts	[29],[30],[31]
		serine-threonine protein kinase PRP4	687-1007	HIV-2	inhibit PRP4-mediated phosphorylation of SF2	[32]
		APOBEC3C	R122[33]	Subtype B	A3C incorporation into virions	[33, 34]
		ATPase KIF4	601-702[35]	HIV-1	Gag intracellular Trafficking	[35],[36]
		virion-associated nuclear shuttling protein(VAN)	312-637	Subtype B	nucleocytoplasmic transport in nucleus	[37]
		eukaryotic initiation factor2α(IF2)	921-1220	HIV-1	inhibits translation	[38]
		Rch1(32AA N-terminally truncated form of Importin α1)	244-529	Subtype B	PIC nuclear import	[39]
Capsid	P85-	Cyclophilin A(CypA)	W121Y,R55,F113[4	HIV-1,	Viral core uncoating &	[48],[40],[49],

	I91[40],T54,A92, R132[41],A92E,G94 D[42], H219[43],P221,P222 [44], N74D[45],P85-A88,A92,P93,G94		6],54-126	FIV, SIV[47]	incorporate into viral budding	[41],[50],[51]
	N53,L56,N57,M66, Q67,K70,I73,N74,A 105,T107,S109,Y13 0	CPSF6 (CFI_m)	V314,L315,F321	HIV-1,HIV-2,SIV,FIV	CPSF6 binds CA in post-entry stages before Uncoating, nuclear import	[52],[53],
	G89[54],H87[55], P38[56], V83, G89, H120, P122,W117, Y130,W133[57]	TRIM5α	Q109,E110,V114,I1 15,L118,E120,R121, Q123,E124[58], R332P[59],W117,L1 18,R121	HIV-1, HIV-2	TRIM5α promotes capsid disassembly during viral uncoating in cytoplasm	[60],[54],[61], [55],[56],[58], [59],[57]
	E45,T54,N57,Q63,Q 67,N74,A105[62]	transportin 3 (TNPO3 or TRN-SR2)	Last 93AA[63]	HIV-1	PIC nuclear import	[62], [63], [64],
	N74	Nucleoporin NUP98		HIV-1	PIC nuclear import	[65],[66],
	N57,Q67,K70,N74[6 7]	Nucleoporin NUP153	896-1475	HIV-1	PIC nuclear import	[68] ,[67], [69]
	G89,P90,I91[70]	Nucleoporin NUP358 (RanBP2)	V61,V113	HIV-1	PIC nuclear import	[70],[71],[65], [72]
	S16,P17	peptidyl prolyl-isomerase PIN1		HIV-1	CA core uncoating in cytoplasm	[73]
	V3	clathrin adaptor complex 2 AP-2, μ2 subunit		HIV-1	Gag intracellular Trafficking	[16]
	V3	adapter protein complex 2 AP-2, α1 subunit		HIV-1	nuclear translocation of viral DNA in cytoplasm or perinuclear region	[74]
	1-110	ankyrin-1, Ank(GAG)1D4		HIV-1	Integration blockage at the post-integration phase	[75]
	177-231	lysyl-tRNA synthetase LysRS		HIV-1	LysRS packed into virions	[76]
		PDZD8	932-1119	Subtype B	After viral entry in cytoplasm	[77]
		actin cross-linking protein filamin A (FLNa)	2364–2647	Subtype B	Gag intracellular Trafficking	[78]
SP1						
Nucleocapsid	K34,C49,N55	Moloney leukemia virus 10 (MOV10)	261–305	HIV-1	MOV10 packaging during virion budding	[79],[80]
	R3,R7,R10,K11,K14 ,K20,R26[81]	ALIX (AIP-1)	F99-A112[82],Q8,K11,K 48,R51,R56,K60[83]	Subtype B	Recruit gag to plasma membrane in viral budding	[81],[84],[85], [82],[83],[85]
	M1-K11[86],R29-K34[87]	APOBEC3G (A3G)	Y124-W128[88]	HIV-1	A3G incorporation to virions in viral budding	[86],[89],[88]
		APOBEC3F(A3F)	W126,R305,Y307	Subtype B	A3F Incorporation	[90]
		APOBEC3A(A3A)	Y130-W133	HIV-1	A3A incorporation	[91]
	Y36-P49	mRNA binding protein 1 (IMP1)	409-458,490-540	HIV-1	Impedes gag assembly, keep immature virus on cellular membranes	[92]
	C15-C49	double-stranded RNA-binding protein Staufen 1 (Stau1)	26-37	HIV-1	Stau1 packed into virions, influences gag multimerization	[93]
	K14,K20,R26,R29,K 33,K34,K38,K41,K4 7	ATP-binding protein ABCE1 (HP68)		HIV-1	capsid assembly	[94]
	43-48	Topoisomerase I(TOP1)		Subtype B	enhancing reverse transcription	[95]
	K14,K20,R26,R29,K 32-K34,K38,K41,D48	Elongation factor 1-alpha (EF1a)	1-74	Subtype B	inhibition of translation	[12]
	R3,R7,R10,K11,K14 ,K20,R26	Nedd4-like ubiquitin ligase, Nedd4-1	L211-L214	Subtype B	Viral release	[85]
SP2						
p6	Y36-L41[96], E34,L35,P37, L41,R42[97]	ALIX (ALG-2 interacting host protein, AIP-1)	F99-A112[82],V498,F67 6,I683[96]	Subtype B	HIV-1 buds via the Alix driven pathway, ALIX incorporates into virions	[81], [98], [85],[82], ,[97]
	P7-P10,	Tumor susceptibility gene 101(TSG101)	391-510	HIV-1	Form viral budding machinery to bud from plasma membrane	[81], [99],
	K27	small ubiquitin-like modifier SUMO-1		Subtype B	ESCRT-III recruitment to viral budding	[100],
	K27	Ubc9		Subtype B	ESCRT-III recruitment to viral budding	[101], [100],
	K27	Daxx		Subtype B	ESCRT-III recruitment to viral budding	[100],

	T23	mitogen-activated protein kinase (MAPK/ERK-2)		Subtype B	P6 phosphorylation within HIV-1 virion	[102]
	L52-H56	Clathrin		HIV-2, SIV	clathrin incorporation	[103]
	L35-L38	Nedd4-like ubiquitin ligase, Nedd4-1	L211-L214	Subtype B	Viral release	[85]
	P5,P7,P10,P11,P24, P30,P37,P49	Cyclophilin A(CypA)		HIV-1	Catalyzes prolyl cis/trans interconversion of p6 Pro residues	[104]
	K27,K33	Ubiquitin		Subtype B	assembly and budding	[105]
gag		tripartite motif containing 22 TRIM22	C15,C18 at RING domain	HIV-1	Disrupt gag trafficking to extracellular membrane	[106]
		tripartite motif containing 21 TRIM21	T14-C59,C97-H129	HIV-1	TRIM21 incorporated to virions	[107]
Protease	32,71	focal adhesion plaque proteins(FAK,), beta 4 integrin,alpha 3 integrin		HIV-1	Affect focal adhesion plaque integrity	[108]
Reverse transcriptase	G462,464-541	AKAP149(A-kinase anchoring protein 149, also AKAP1)	375-645	Subtype B	reverse transcription	[109]
	1-243	APOBEC3G (A3G)	65-132	HIV-1	A3G inhibits reverse transcription.	[110]
Integrase	A128,A129,W131, W132,I161-K173, T174,T125,	LEDGF/p75	I365,D366,F406,K415,V408[111]	HIV-1, HIV-2, FIV	IN strand transfer activity	[112],[113],[114], [115]
	185-188,R262-K264,266-269[116]	Transportin-SR2 (TRN-SR2,Transportin 3,TNPO3, importin 12)	62-334[117], R400,R402,Q761[118]	HIV-1	PIC nuclear import	[116],[117],[118],[119]
	161-173[120], K186-K189,K211-K219[121]	Importin α1 (karyopherin α2, Rch1/SRP1α/KPNA2)		HIV-1	PIC nuclear import	[121], [120], [119], [122]
	251-270	Importin α3		Subtype B	PIC nuclear import	[123]
	K186-K189[122],	Importin β1		Subtype B	PIC nuclear import	[122]
	W235-G245,R262-K266[124]	importin 7 (IPO7)		HIV-1	DNA genome nuclear import	[124],[125]
	H12,K71,Q137,S147,D202[126],E69G,K71R[127]	INI1(hSNF5, BAF47, SMARCB1)	T213,D224,D226,S246	HIV-1	Gag/gagpol trafficking to membrane, IN multimerization and integration	[126],[128],[129],[130],[131],
	K264,K266,K273	Histone acetyltransferase p300/CBP		Subtype B	lysine acetylation	[132],[133]
	Y15,V75,A76,K186, L241,L242[134],	survival motor neuron (SMN)-interacting protein 1 (SIP1/Gemin2)	137-238	Subtype B	stimulate RT activity	[135],[134],
	43-195	von Hippel-Lindau binding protein 1(VBP1)		HIV-1	integrase degradation in integration-transcription transition	[136]
	50-211(CCD)	Huwe1	1-3617	HIV-1	localization of gagpol precursor	[137]
	212-264[138]	embryonic ectoderm development (EED) protein	L96-H104,T224-V232[8]	HIV-1	DNA integration in the nucleus and near nuclear pores	[138],[8]
	48-212	Heat shock protein HSP60		HIV-1	Stimulate IN activity within PIC	[139]
	L172,170-180[140]	Uracil DNA glycosylase (UNG) 2	1-52[141]	Subtype B	UNG packed into new virions	[140],[141]
	K264, K266, K273	KAP1(TRIM28, Tif-1 β)		Subtype B	KAP1 Inhibits HIV-1 Integration	[142]
	K258, K264, K266, K273	transcriptional activator GCN5		Subtype B	acetylates IN leading to enhanced 3'-end and strand transfer activities	[143],
	230-288	Ku70(XRCC6)	1-430	HIV-1	Ku70 protects IN from proteasomal degradation	[144]
	N184,K185[145]	clathrin		HIV-1	clathrin incorporation	[103],[145]
	Y23	peptidyl prolyl-isomerase PIN1		Subtype B	controls IN stability during DNA integration	[146]
	K46,K136,K244	small ubiquitin-like modifier SUMO-1		HIV-1	IN SUMOylation, proviral integration	[147]
	K156-K159[148]	Daxx	625-740[149]	HIV-1	Viral DNA integration	[149],[148]
	E87,V88,I89,P90,E96,Y99,F100,K103,K173	Sucrose		HIV-1		[150]
	50-212	APOBEC3G (A3G)	104-156	HIV-1	Reduce reverse transcription of proviral DNA	[151]

		HRP2(hepatoma-derived growth factor related protein 2)	470-552[114]	HIV-1	IN strand transfer activity	[114],[152],[153]
		Rad18 (RNF73)	65-226	HIV-1	integrase stabilization	[154]
		TTRAP(tyrosyl-DNA phosphodiesterase 2)	146-362	HIV-1	integration	[155]
		Nucleoporin NUP153	896-1475[156]	HIV-1	Integrase nuclear import	[156]
		Nucleoporin 62	328-522	HIV-1	Integrase nuclear import	[157]
Vif	W5,W11,D14-R17,W21,L24,V25, W38,L64,I66,Y69, W70,L72,W79,W89 [158], S23-V25[159]	APOBEC3C,	E106-H111	HIV-1	APOBEC3C degradation	[158],[160],[161], [159]
	S23-V25[159], W5,W11,D14-R17,W21,L24,V25, W38,L64,I66,Y69, W70,L72,W79,W89 [158]	APOBEC3D(APOBEC3DE)	L268,F271,C272,I275,L276,S277,Y282, H307, E302,F303, E337	HIV-1	APOBEC3D degradation	[158], [159],[159]
	K26,H27[162],K22, Y40,E45[163],W11, Q12,D14-R17,V25,Y69-L72, T74-W79,L81-G84,E171-W174,I87,W89[164]	APOBEC3F	L255,F258,C259,I262,L263,S264,Y269, E289,F290, H294, E324[161]	HIV-1	APOBEC3F degradation	[160],[161],[165],[166],[163],[167]
	W5,W21,K22,K26, Y30,W38,Y40-Y44,L59,L64,I66,Y69,W70,L72,W89[158],S23-V25[159],N48[168], G84[164],L81,G82,I87,W89[164]	APOBEC3G (CEM15)	D128-D130[88], [169]	HIV-1, HIV-2	APOBEC3G degradation	[160], [166], [170],[171],[172],[168],[169],[164],[167]
	D14,R15,F39,R41,H42,H48,L64,I66,Y69 ,L72[158]	APOBEC3H	E121[173]	HIV-1	APOBEC3H degradation	[158]
	L145,Q146,A149,L150[174],I120,A123, L124[175],H108,C114,C133,H139,159-173,P161-164,L169	Cullin5 of E3 ubiquitin ligase(Cul5)	Y75,F93,L103,A107 ,C112[174]	Subtype B	Vif-mediated downregulation of APOBEC3	[160],[172],[175],[176] ,[177]
	4-22	MDM2 E3 ubiquitin ligase(Hdm2)	168–320	Subtype B	MDM2 reduces cellular Vif levels and reversely increases A3G levels	[178]
	20-128	E3 ubiquitin ligase Nedd4,AIP4	WW domains	Subtype B	Vif ubiquitination	[179]
	P161-P164 [180]	hemopoietic cell kinase HCK	66,68,75,93[180]	Subtype B	Vif represses the kinase activity of Hck	[180],[181]
	P161-P164 [177]	Elongin B	D101-K104[177],9-14 [182]	Subtype B	Vif-mediated downregulation of APOBEC3	[160],[177], [182]
	S144-Q146,P161-P164 [177],V142,L145,L148,A149,A152,L153 ,L163,V166,L169,R173[176], L158, C162,V165,V166[183]	Elongin C	Y79,Y76,V73,I90,L103,L104,N108,	Subtype B	Vif-mediated downregulation of APOBEC3	[176], [177], [182], [184],[183]
	96,165	mitogen-activated protein kinase (MAPK)		Subtype B	MAPK phosphorylates and regulates Vif	[185]
	4-22	tumor suppressor protein p53 (TP53)	168-320	Subtype B	Vif-mediated G2 cell cycle arrest	[186]
	W21,W38[187]	CBF-β, core binding factor β	F68[188],D101-K104[182],69-90,129-140[187]	Subtype B	Essential for assembly of Vif–CUL5 E3-ubiquitin-ligase complex	[182],[187],[188]
	12-23,43-59,73-87,97-112	nuclear body protein Sp140	527-836	Subtype B	Vif may enhance cytosolic retention of Sp140	[189]
Rev	35-50,75-84	LEDGF/p75	361-370,402-411	Subtype B	promote dissociation of IN-LEDGF/p75 complex	[190]
	L78,L81[191],L75-L83[192]	RIP (Rev-interacting protein, REBP, Rev/Rex effector binding protein, Rab)	148-271 (FG repeats) [193],	HIV-1	Rev nuclear export	[194],[193], [195], [192],[191]
	9-14	heterogeneous nuclear ribonucleoproteins A1(hnRNP A1)	R194,R196,R206,R228,R235R242	HIV-1	Multiple functions	[196]
	L78,E79[197]	Sam68 (Src-associated protein in mitosis)	321-410[197]	HIV-1	Rev nuclear export	[197],[198]
	18-50	Puralpha	73-123	HIV-1	enhance Rev-RRE binding in cytoplasm	[199]
	L75-L83	Chromosome maintenance region 1 (CRM1,exportin 1, XPO1)	D716	HIV-1	nuclear export to nuclear pore complex	[200]

	10–24[201],V16,I55[201]	RNA helicase DDX1	370-373	HIV-1	Promote Rev oligomerization	[202],[203],[204],[205],[201]
	59-116[206]	Nucleoporin Nup98	493-920[206]	Subtype B	Rev nuclear export	[206], ,[207]
	L75-L83	Nucleoporin Nup214(CAN)	1691-1894[195]	Subtype B	Rev nuclear export	[208], [195]
	35-46[209]	Importin beta (karyopherin β)	1-396[210]	HIV-1	nuclear import into nuclei	[210],[211],[209]
	R38-R44	Importin 5		HIV-1	Rev nuclear import	[210]
	R38-R44	importin 7 (IPO7)		HIV-1	Rev nuclear import	[210]
	1-59	microtubules		HIV-1	Microtubule destabilization	[212]
	35-46	HIC (Human I-mfa domain-Containing protein)	144-246	HIV-1	Rev nuclear import	[213]
	R41-Q51	NAP1(nucleosome assembly protein 1)		HIV-1	Nap1 affects Rev multimerization & nuclear export	[214]
	37-47[215]	Nucleolar phosphoprotein B23(NPM1, NO38, numatrin)	187-255[216]	HIV-1	Rev nuclear export in nucleus	[217],[216],[215]
	34-50	p32(splicing factor ASF/SF2-associated protein)	196-208	HIV-1	mediate Rev activity in RNA splicing in nucleus	[218]
	S5,S8[219],35-50[220]	protein kinase CK2 beta		HIV-1	CK2 phosphorylate Rev	[220],[219]
	73-84	Nuclear prothymosin alpha ProTa		HIV-1	Rev nuclear import	[221]
	75-83	NLP-1 (nucleoporin-like protein 1, hCG1,NUPL2)	213-380	HIV-1	Rev nuclear export	[222]
	75-93[223]	eIF-5A (eukaryotic initiation factor 5A)		HIV-1	Rev nuclear export	[224],[225],[223]
		HS1-associated protein X-1 (Hax-1)	176–260	HIV-1	inhibits Rev from binding to RRE RNA	[226]
		mRNA binding protein 1 (IMP1)	1-408	HIV-1	modulate RNA expression, relocate Rev from nucleus to cytoplasm	[227]
		Transportin 1	1-517	HIV-1	Rev nuclear export	[210],
		RNA helicase DDX5 (p68)	248-251	Subtype B	Rev nuclear export	[228],[204],[205]
		RNA helicase DDX3,DDX17,DDX21,DDX56	11-21	HIV-1	Rev nuclear export	[204],[205],
Vpu	14,22,18[229],E15,V19,V25 (group N)[1],	Tetherin(BST-2, CD317, HM1.24)	22-46[230],9,17,43,52[231], 34,37,41,45[232]	Group M[2], group N,SIV[1]	Vpu-mediated ubiquitination degrades BST-2 in the trans-Golgi network or in early endosomes	[229],[230],[232] ,[1]
	K31-Q35[233]	SGT(glutamine-rich tetratricopeptide repeat protein, also vpu-binding protein UBP, VIP)	84-210	Subtype B	Viral release	[234],[233]
	L41,S52,S56,E57,D79,D80	CD4	414-419	Subtype B	CD4 degradation in the endoplasmic reticulum	[235], [236]
	S52,S56	CK-2(casein kinase 2)		Subtype B	Vpu phosphorylation	[237],[238]
	L45,I46,S52,G53,S56[239], S61 [240],	Beta-transducin repeat-containing protein (Beta-TrCP)	260-293	Subtype B Subtype C	inhibit p53 ubiquitination, proteasomal degradation, CD4 degradation	[235], [241],[239, 240], [240]
	S52,S56	Interferon regulatory factor 3 (IRF3)		Subtype B	Vpu redirects IRF3 to endolysosome for proteolytic degradation	[242]
		wik-related Acid Sensitive TASK-1	1-40	HIV-1	transcription	[243],[244]
Vpr	17-46[245],L23,K27,A30,F34[246]	nucleoporin CG1 (hCG1, NLP-1)	94-170[245]	HIV-1	Vpr interacts with NPC components at nuclear envelope	[245],[246]
	17-34,46-74[247]	Importin α1 (karyopherin α2, Rch1/SRP1α/KPNA2)	393-462[247], 404-475[248]	HIV-1	Vpr enters nucleus by interacting with nucleoporins at nuclear pore	[249] ,[250], [247],
	17-34[250]	Importin α3(Qip1,karyopherin α4	392-439 [248]	HIV-1	PIC nuclear import	[248], [250]
	17-34[250]	Importin α5(NPI1,karyopherin α1	404-451 [248]	HIV-1	PIC nuclear import	[248], [250], [247],
	F72[251]	Importin 5 (importin β3)		HIV-1	PIC nuclear import	[251],
	17-34[250]	Importin β (karyopherin β)	71-876[252]	HIV-1	PIC nuclear import	[252],[251],[250],
	65-85	Transcriptional coactivator p300	2045-2191		activate HIV transcription	[253]
	W54,E25,N16-W18[254], W54R,S79A	uracil DNA glycosylase (UNG2)	W222-W225[255]	HIV-1	induce G2 arrest by proteasomal and polyubiquitination degradation in cytoplasm	[254],[255],[256]
	Q65R,[257],L60-81[258]	DCAF1(VprBP)	877-1365[259]	Subtype B	Vpr-induced G2 arrest	[260],[257],[259],[258]

	71-96	HS1-associated protein X-1 (HAX-1)	118-141	Subtype B	mitochondrion instability, Vpr proapoptotic activity	[261]
	71-82	adenine nucleotide translocator (ANT, component of permeability transition pore complex PTPC)	104-116	HIV-1	modulate mitochondrial membrane permeabilization (MMP)	[262], [263]
	I74,I81	Wee1	291-369	HIV-1	Vpr-induced G2 arrest	[264]
	E25,H33,H45,G75[265]	HHR23A(RAD23A)	319-363[266],I324,A329,L330,F332,P333,A353,F354,L356,F360[267]	Subtype B	Involve cellular DNA repair pathway to induce life cycle G2 arrest at or about the nuclear membrane	[265],[266],[267]
	1-35[268],15-77[269],R80[270]	basal transcription factor TFIIB	W52,F55[270]	HIV-1	stimulate transactivation activity in nuclei	[268],[208],[270]
	L67,R80,R88	SAP145	426-563[271]	HIV-1	Vpr-induced checkpoint activation and G2 arrest	[271],[272]
	H71,R90	CDC25c	E352,K359	HIV-1	inactivation of the cdc2-cyclin B kinase complex, promotes G2 arrest	[273]
	R80	14-3-3tau	190-210	Subtype B	promotes G2 arrest	[274]
	L22-L26,L64-L68	Chromosome maintenance region 1 (CRM1,exportin 1, XPO1)		Subtype B	Vpr nuclear export nucleocytoplasmic shuttling	[275]
	L64	damaged DNA-binding protein (DDB1)		Subtype B	Vpr-induced G2 arrest	[276],[257]
	L64-L68[268],L22-L26,	glucocorticoid receptor (GR)		HIV-1	Activate glucocorticoid signal transduction pathway	[268],[277],[278]
	H71,G75	replication protein A (RPA)		HIV-1	chromatin binding activity, induces ATR activation	[279]
	1-39	Lys-tRNA synthetase(LysRS)		Subtype B	inhibits LysRS-mediated aminoacylation to affect initiation of reverse transcription	[280]
	Q60-P81	transcription factor SP1		HIV-1	Vpr trans-activation of HIV-1 LTR promoter	[281],
	H71-G75	Chromatin-remodeling factor SNF2h		HIV-1	downregulate endogenous to induce SNF2h DNA double-strand breaks	[282]
	77-92[283]	Protein phosphatase 2A1 (PP2A)		HIV-1	life cycle G2 arrest in cytoplasm	[283],[284]
		cyclin T1	300-479	Subtype B	Viral transcription	[285]
		Nucleoporin Pom121	796-1199	Subtype B	PIC nuclear import through NPC at nuclear envelope	[286]
		Glucocorticoid receptor p21 (WAF1, Cip1)	1-90,149-164	Subtype B	Vpr alleviates p21-mediated inhibition of cell departure from G1 phase	[287]
		VIP/mov34	225-341	HIV-1	Vpr-induced G2 arrest	[288],[289]
Vpx	Q76[290],N12,E15,E16,T17[291],	SAMHD1	595-626[292]	HIV-2, SIV	Vpx induces proteolytic degradation of SAMHD1	[290],[292],[291],[293]
	101–112	alpha-actinin 1	346–892	HIV-2, SIV	Vpx and PIC nuclear import	[294]
	61-112	heat shock protein 40, Hsp40/DnaJB6	100-326	HIV-2	PIC nuclear import	[295]
	Q76,K77[296], W24,K68,K77[297], V29,I32,A36,V37,H39,Q76,F80[293]	E3 ubiquitin ligase DCAF1		HIV-2	Macrophage infection	[296],[297],[293]
	H82	APOBEC3A		HIV-2	Vpx counteracts APOBEC3A in cytoplasm	[298]
	103-106	Src-like tyrosine kinase Fyn SH3		HIV-2	PIC nuclear import	[299]
	Y66, Y69, Y71[299],	mitogen-activated protein kinase (MAPK/ERK-2)		HIV-2, SIV	Vpx phosphorylation	[299],[300]
		CD74(MHC II invariant chain, Ii)	134-216	HIV-2	Vpx may disrupt major histocompatibility complex class IIantigen presentation	[301]
Tat	24-36,45-72[302], 49-86[303],52-57[304]	RNA polymerase IIa	1325-1630[304]	HIV-1, HIV-2[304]	induce CTD phosphorylation and transcription from HIV-1 promoter	[302],[305], [303],[304], [306],[304]
	P18,C22,K41[307],3,6,10,47	cyclin T1 (CCNT1,a component of PTEFb, a heterodimer of CycT1 and CDK9) (CDK9=PITALRE=TAK)	44,111,112,253–256	HIV-1, HIV-2[308]	transcription elongation to TAR RNA	[302],[307], [309], [310], [311],[312],[308]
	K28[313],K50,K51[314],20-	Histone acetyltransferase	1542-2412[317],F748,V75	HIV-1	Tat acetylation, transcriptional activation	[317],[318],[321],

	40[315],Y47,R53[316]	p300/CBP-associated factor (P/CAF)	2,Y802,Y809[318], 1253-1710[319],760,761[320], V763,Y802,Y809[316]			[322],[313], [323],[314],[315],[316]
	47-57[324]	tumor suppressor protein p53(TP53)	K351[325],E343, E349, 326-355[324]	HIV-1	Tat-induced transactivation	[325],[324]
	K51,R52,D67	20S proteasome	E235,K236,K239	HIV-1	Affect proteasome function(antigen processing	[326]
	R49-R57[327]	IkappaB-α	263-269[328]	HIV-1	NF-kappaB deregulation	[327],[328]
	K50,K51[329], 32-48[330]	splicing regulator p32(32-kDa protein, TAP)	244-260[331]	HIV-1	Tat acetylation affected splicing of HIV-1 genome	[332],[329], [331],[330]
	P3-P6, P81-P84	Grb2 (growth factor receptor-bound protein 2)	160-212	HIV-1	Inhibit Tat-mediated transactivation	[333]
	1-49[334]	transcription factor E2F-4	1-184	HIV-1	transcriptional activity	,[334]
	K41,K50,K51[335]	human sirtuin 1 SIRT1	F414,V445,P447[335]	HIV-1	Tat inhibits SIRT1 deacetylase activity	[335],[336]
	21-47	HT2A (TRIM32)	526-653	HIV-1, HIV-2	transactivation	[337]
	72,86	INI1(hSNF5, BAF47,SMARCB1)	Rpt1(183-294) and Rpt2 domain	HIV-1	transcriptional activation	[338]
	K41,K50,K51	BRG-1 (SWI/SNF chromatin-remodeling complex)	1400-1700[339]	HIV-1	transcriptional activation	[339],[340]
	37-72	microtubule-associated protein LIS1 (a subunit of platelet-activating factor acetyl hydrolase)	WD domain 5	HIV-1	Tat-induced apoptosis to distortion of microtubules polymerization	[341]
	67-101[342], 18-36,36-56[343]	TAFII250 (TAF1 RNA polymerase II)	848-1279,885-984,1120-1279	Subtype B	Tat repression of MHC class I transcription	[342],[343]
	49-86	mammalian mRNA capping enzyme Mce1	211-597	HIV-1	stimulates capping of TAR RNA during transcription	[344]
	22-48	histone acetyltransferase GCN5	111-251,389-476	Subtype B	transactivation	[345]
	49-72[346]	Puralpha	100-121[346]	HIV-1	transcriptional activation[346], deregulation of NGF transduction pathway[347]	[346],[348],[347]
	48-72	Y-box protein YB-1	75-203	HIV-1	Tat-induced transactivation	[349]
	30-72	Dicer (ATP-dependent RNase III)	246-585	HIV-1	Tat-induced transactivation	[350]
	1-48	chicken ovalbumin upstream promoter transcription factor-interacting protein 2 (CTIP2)	145-434	Subtype B	Tat inactivation through subnuclear relocalization	[351]
	1-27[352]	NFAT1 (nuclear factor of activated T cells)	1-96[352]	HIV-1	NFAT1 inhibits Tat-mediated activation of HIV-1 LTR transcription	[352],[353]
	R49-R57	NAP1(nucleosome assembly protein 1)	162-290,290-391	HIV-1	Tat-induced transactivation	[354]
	R49-R57	Nucleolar phosphoprotein B23(NPM1)	187-255		Rev nuclear export	,[216]
	73-101(second exon)	human translation elongation factor 1-delta (EF-1δ)	144-280	Subtype B	Repress RNA translation	[355]
	Q35-F38, V36,F38	serine-threonine protein phosphatase PP1γ	RVXF motifs	Subtype B	stimulates Tat-mediated transactivation of HIV-1 LTR promoter in nucleus	[356]
	1-40	p160 coactivator GRIP1	1-97	HIV-1	HIV-1 LTR Transactivation by Tat	[357]
	C30,C31[358],31-61[359]	N-methyl-D-aspartate receptors NMDAR	C744[358]	Subtype B	neuropathogenesis	[359],[358]
	36-50	TATA-binding protein (TBP) subunit of TFIID	last 181 amino acids	HIV-1	transcriptional activity	[360],[361]
	30,41,10-48	transcription factor TFIIB	27-103,148-163	HIV-1	Stabilize transcriptional initiation complex	[362]
	1-48	CDK7 in TFIIH kinase		HIV-1	Tat transactivation	[363],[364]
	K13, L18, K21, T26, I27, T29, G33, L39, A57	calmodulin		HIV-1	control multiple metabolic pathways.	[365]
	1-45	Toll-like receptor 4 TLR4-MD2		HIV-1		[366]
	C22,C25,C27[327]	Transcriptional coactivator p65		HIV-1	NF-kappaB deregulation	[327],
	K50	Endogenous BRM(a DNA-dependent ATPase subunit of SWI/SNF)		HIV-1	transactivation	[367]
	R49,R52,R53,R55,R56,R57,R78-D80[368]	αvβ3 integrins		HIV-1	activate focal adhesion kinase in extracellular	[368], [369]
	G48-R57	α5β1 integrins		HIV-1	Activate adhesion kinase in endothelial cells	[369]
	45-86[370]	α4β7 integrin		HIV-1	cell attachment to tat in extracellular	[370],
	22,30	TIP60 (Tat interactive protein, 60 kDa)		HIV-1	transactivation	[371]
	15-24,36-49	CDK2/cyclin E		HIV-1	Cdk2 phosphorylates	[302]

					tat,transcription	
	1-48[372]	TIP30(CC3,SDR44U1,HTATIP2)		HIV-1	transactivation	[372],[373]
	47-67	CCAAT/enhancer binding protein(C/EBPβ)		HIV-1	activation of MCP-1 transcription	[374]
	46-60	VEGF receptor Flk1/KDR		HIV-1	Tat-induced angiogenesis	[375]
	24-51	β-chemokine receptors CCR2		HIV-1	Receptor activation in extracellular	[376]
	24-51	β-chemokine receptors CCR3		HIV-1	Receptor activation in extracellular	[376]
	24-51	CXCR4		Subtype B	viral entry interference	[377]
	37-48	lipoprotein related protein receptor (LRP)		Subtype B	neuropathogenesis	[378]
	36-39	microtubule αβ-tubulin		HIV-1	Tat-induced apoptosis	[379]
	R49-R57	Importin β		HIV-1	Tat nuclear import into nuclei	[209]
	49-57	heparan sulfate proteoglycans (HSPG) of syndecan-2, syndecan-4, and CD44v3		HIV-1	Tat transduction in extracellular	[380],[381]
	R49-R57	vascular endothelial growth factor receptor type 2 (VEGFR2,KDR)		HIV-1	endothelial cell activation in extracellular	[382]
	F31-G60	dextrin-2-sulfate(D2S)		HIV-1	D2S inhibits tat transactivation	[383]
	C22	dopamine transporter (DAT)		HIV-1	Tat regulates DAT activity on plasma membrane	[384]
	48-60[385],46[386]	protein kinase C- α(PKC- alpha)		HIV-1	inhibit PKC phosphorylation of ERK1/2 in cytoplasm	[385],[386]
	40-58[387]	RNA-activated kinase PKR		Subtype B	Phosphorylation of Tat by PKR, Tat inhibits PKR autophosphorylation	[387],[388]
	K71	MDM2 E3 ubiquitin ligase (Hdm2)		Subtype B	Hdm2 ubiquitinates Tat	[389]
	R53-R56	furin		HIV-1	furin cleaves tat to inactivate extracellular tat	[390]
	1–22, 38–53, 93–101	MAb 7G12				[391]
	E9-K12	Fab 11H6H1 antibody	G91,Y58	HIV-1		[392]
		Werner syndrome helicase (WRN)	K577	HIV-1	Recruit transcription complex	[393]
		DDX3	536-661	HIV-1	transactivation	[394],[395]
GP120	101-120,160-175,252-261,308-328	protein-disulfide isomerase (PDI)	acidic C terminus	Subtype B	env biosynthesis in ER	[396]
	317,318,320,322	Salivary agglutinin gp340 (SAG)	33-36,73-76,94-97	Subtype B	gp340 binds GP120 to inhibit coreceptor binding	[397],[398],[399]
	365-371,425-430,112,255,256,257,368-371,375-377,382,384,427,473,475	CD4	Q40P, F43L,G47R [400],K29,K35,K46, R59,F43[401]	HIV-1	GP120-receptor binding	[402],[403],[404],[405],[406],[407],[400].[401]
	296-331(V3 domain)	CCR5	2-15[404]	HIV-1, HIV-2, SIV	GP120-receptor binding	[408],[409],[404]
	296-331(V3 domain)	CXCR4	E14,E15,D20,Y21,D22,D187-Y190,D193,D262, E268,E277,E282[408]	HIV-1, HIV-2, SIV	GP120-receptor binding	[408],[409],[410],[411]
	296-331(V3 domain)	CCR3		HIV-1	GP120-receptor binding	[412],[410],[413]
	301-317	immunophilins FK506-binding protein (FKBP12)		Subtype B	immunophilins binds gp120 in extracellular	[414]
	31-50,101-120,160-175,308-324	Calnexin(CNX)		Subtype B	env biosynthesis in ER	[396]
	R508-R511[415]	furin		HIV-1	Cleaves gp160 into gp120 and gp41 in ER	[415],[416],[417]
	R419,K421,K432	heparan sulfate proteoglycans (HSPG,SDC2)		Subtype B	co-receptor recognition in extracellular	[418]
	204,259,309	mannose-binding lectin Concanavalin A (ConA)		Subtype B		[419]
	230,289,295,386,392,448[420],234,241,289,339[421],262,332[422]	Lectin griffithsin(GRFT)	30,70,112[423]	A,B,C		[420],[424],[421],[422],[423],[425]
	230,234, 289, 295, 332, 339, 386,392,448[424],397[426],302,362,367, 376,418[419]	mannose-binding lectin Cyanovirin-N	2-3,7,23-27,93-95,41-44,50,53,56,57,74-78,93	A,B,C		[427], [424],[419], [428],[429],[430], [431]
	332,392,339,295	mAb 2G12	19		Inhibit GP120-receptor binding[432]	[433],[434], [435],[436],[437],[432]
	197,301,364,369,372	mAb b12				[402],[438],[4

	,373[407]					39],[440],[407]
	334,386,392,397,450	mAb 17b[441]				[442],[439],[443],[440]
	S256,T257,D368-E370,K421,P470-484[444]	mAb F105		Subtype B		[445],[440],[444]
	420–423	mAb E51				[443],[446],[447]
	298,302,303,441	mAv 412d				[404], [443]
	176,177,179,180,183,184,192-194	mAb 697-D		Subtype B		[448]
		mAb 830A				
	510-516	mAb 1331A				[441]
		mAb VH1-69				
	230,234,276	mAb 8ANC195				[449]
	293,334,337,340,460	mAb 19.3H-L1,L3	295,333,335			[450]
	119,199,207,298,366,367,368,392,423,426,427,430,432,435,437,472,473	mAb m6				[451]
	227,229,233	mAb m9				[451]
	227,233,234,423,432	mAb scFv X5				[451]
	D368	mAb m43				[452]
	295,332,392,386,448[453]	mAb 2G12		A,B,C		[440], [453],
	F159,N160,L165-D167,K169,K171	mAb CAP256		A,B,C		[454]
	309,312-317	F425-B4e8 (B4e8)	32,92,100	B,C,D		[455],[456]
	K160,295,406,448,463	3BC176, 3BC315		A,B,C		[457],
	156,158-160,162,173,176,181,299	PG9	96-102	A,B,C,D,G,F,01_AE,02_AG		[458],[459],[460],[461]
	156,158-160,162,173,176,181,299,305,307,309,317,318	PG16	96-102	A,B,C,D,G,F,01_AE,02_AG		[458],[461],[462]
	301,332	PGT128				[463],[464]
	127,159,160,168,169,171,181	CH01, CH02, CH03, CH04				[465]
	D474,R476,M475,R476[466]	HJ16				[466],[467],
	307-312,315-317	HGN194		Subtype C		[468],[469]
	50,58,67,71,96,100	VRC01	276,279,280,456,459,368			[402],[470],[471], [472]
	S199,N276,D279,N280,I420,I423,D457,N461,S463,R469	PGV04(VRC-PG04）				[473],[474]
	K305,H308,R315	Fab 268-D				[475]
	308,309,312-318	Fab 1006	Y32,W91(light chain),D31,W33,Y52,P53,D54,D56(heavy chain).	A,B,C		[475], [476]
	307-318	Fab 2219	Y32,W91(light),D31,W33,Y52,P53,D54,D56(heavy)	A,B,C		[475],[476],[477]
	304-309,312-318	Fab 2557	K31,Y32,W91-L98(light), D31,W33,Y52,P53,D54,D56,H58,L95,L97,N100(heavy)	A,B,C		[475], [476]
	307,309,313,317	Fab 3074	Y49(light),S30,Y53,F96,E98-Y100(heavy)	A,B,C		, [476]
	160,167,169,313,315	Fab 2909	95-100	SF162		[461],[478],[479]
	R304,K305,I307-I309,F317,Y318	Fab 2558	N30,K31,Y32,W91(light),D31,W33,D54,D56(heavy chain)	02_AG		[476]
	R304,K305,I307-I309,F317,Y318	Fab 4025	N30,K31,Y32,W91(light),D31,W33,D54,D56(heavy chain)			[476]
	305-309,312-315	Fab 447-52D	W33,R50,K52,S100-Y106(heavy chain),Y34,W93,A99,W101(light chain)	A,B,C		[480],[481],[482],[483],[484]
	304-309,312-316	Fab 537-10D	W33,W47,N50,Y61,E91,Y95,D97,L99-M108(heavy chain),N30,G31,Y34,Y93,P99-V101	A,B,C		[482]

	312-315	Fab 0.5β	D30,Y32,Y49,E55,F96 (light), H52, Y53, D55,D56,E61,Y96,Y100(heavy)	Subtype B		[485],[486]
GP41	W597–W 611	CD74	72-232	Subtype B	activation of ERK/MAPK pathway during viral entry, extracellular membrane	[487]
	790-811	P155-RhoGEF	860-913	Subtype B	gp41C inhibited p115 mediated actin stress fiber formation and SRF activation	[488]
	751-768	human bZIP transcription factor Luman	100-314	HIV-1	TMgp41 inhibits transcriptional activation mediated by Luman fusion	[489]
	822-855	α-catenin	787-813	HIV-1	GP41 cytoplasmic domain binds to actin filaments	[490]
	670–677	Z13e1	30-33,50-58,97-100	HIV-1	GP41 neutralizing antibody	[491],[492],
	769–788, 826–854(LLP-1)[493],835,838[494]	ubiquitous calcium-sensing calmodulin (CaM, CAMI)		Subtype B	Fas-mediated apoptosis	[493],[495],[494]
	W623-W631	Caveolin-1		Subtype B	Env trafficking to membrane	[496]
	539-684	HSP60		HIV-1	?	[497]
	E633–S650, D675–K685	Retrocyclin-1		Subtype B	retrocyclin-1 prevented 6-helix bundle formation	[498]
	618–623	gC1qR, receptor on CD4+ lymphocytes		HIV-1	induces NKp44L cell-surface expression	[499]
	708-750	major histocompatibility complex class II MHC-II		Subtype B	incorporation of HLA class II proteins	[500]
	650-685	galactosyl ceramide (GalCer)				[501]
	762-773	TGF-beta-activated kinase 1 TAK1		HIV-1	gp41CD-mediated NF-κB activation	[502]
	Y707-L710,Y763-L766[503], Y712, Y721,L855-L856[504], Y712-Y715[505]	clathrin adaptor complex 1 AP-1, μ1 subunit (AP47)		HIV-1, SIV,HTLV-1	Env trafficking	[503], [504], [505],
	Y712-Y715[505]	clathrin adaptor complex 1 AP-1, μ2 subunit		HIV-1	Env trafficking	[505]
	(HIV-1)Y707-L710,Y763-L766[503], Y712-Y715[505], G711-715,L855,L856[506], Y712-L715,Y768-L771[507], (HIV-2)G706,Y707[508]	clathrin adaptor complex 2 AP-2, μ1 subunit (AP50,mu2)		HIV-1,HIV-2	Env trafficking	[503], [505], [506], [507], [508]
	Y707-L710,Y763-L766[503], 712-715[504]	clathrin adaptor complex 3 AP-3, μ1 subunit (mu3A adaptin)		HIV-1	Env trafficking	[503], [504]
	745-751	monoclonal antibody(mAb)SAR1				[509],[510]
	662-670	mAb 2F5	F100B[511]			[511],[512],[513],[514],[515],[516],[517], [518],[519],[491],[520]
	647-682	mAb 1281				[441]
	567-647	mAb 1367				[441]
	735-752	mAb 1575				[28],[510]
	746-750	mAb 1577, MAb 1583				[521]
	671-676[522], 679,680[523]	mAb 4E10	31-33,47,50-58,91-95,100	A,B,C,D, G,F,01_AE,02_AG		[512],[514],[515],[524],[491],[520]
	N671,W672,F673,T676,L679,R683	mAb 10E8				[525]
	672-676	mAb CAP206-CH12				[526]
	664-666,669,671	mAb 13H11	94-96,101			[527]
	564,567,568,571,573-575,577,579, 636,643	mAb HK20	I53-N58[528]			[468] ,[528]
	M593,G594,G597,G600,L602,W610,V612,W614,K617,V619-W623,M626-D636	mAb m44				[529]
	569,675	mAb b12		B,C		[407]
	Q563,H564,Q567,L568,W571,I573,K574,Q575,Q577[528],	mAb D5	I52-N58[528]			[530],[528]

	H564,L568,V570-Q577,D632,N636,T639,H643[530]					
	E560,H564,L565,Q567,L568,V570-Q577,R579,M629,D632,N636,S640,H643,I646,E647	Fab 8066	P27,E30,Y31,Y48, N52,S92,M93,V95			[530]
	E560,Q563,H564,L565,Q567,L568,V570-Q577, R579, N636, H643, I646	Fab 8062	E30,Y31			[530]
	643-661	Fab 1281				[531]
	E560,H564,W571,K574,Q575	Fab 3674				[532]
	786~856	lipid rafts				[533],[534]
	727-732	Chessie 8				[535]
	741-751	EPES		Subtype B	GP41 neutralizing antibody	[536]
	L799,L800	Prohibitin Phb1/Phb2 heterodimer		HIV-1	Phb-dependent cellular process	[537]
	750-763,764-785	Golgi retrieval signals				[538]
	586–596	receptor protein P45		HIV-1	Gp41 binds receptor proteins during viral entry	[539],[540]
Nef	P72-P75[541],76,R77,86,90, 106,109,113,120	Src family kinase Fyn	R96,119,94-100	Subtype B	suppress Fyn kinase and transforming activities, altered T cell receptor signaling	[542],[543] ,[544],[545] , [541], [546]
	P72-P75	Src family kinase c-Src	SH3 motifs	HIV-1	actiate c-Src activity in transGolgi network	[541]
	E62-E65[547],P72-P75,96[548], P72,V74,P75,R77[549],	Src family kinase Hck SH3	D67,E69[548],72-256[550]	HIV-1	actiate Hck activity in transGolgi network	[548], [550], [549], [541], [546], [547]
	P72-P75[541],P69-A83[551]	Src family kinase LCK	SH3 motifs	Subtype B	Nef decreases vitro kinase activity of Lck	[551], [541]
	1~30	N-myristoyl transfterase2(NMT2)	40-50,170-185	HIV-1	Nef is myristoylated by NMT2	[552],[553]
	109,110,	p21-associated kinase 1(PAK1, NAK)	83-149[554]	Subtype B	activation of a PAK-related kinase	[554],
	57-59,60,95,96,97,106, 109,110[555],123[549],L164,L165[556]	CD4	408-418	HIV-1	Nef downregulates CD4	[555], [549], [556]
	57-59	CD28	191,185,193	HIV-1, SIV	Nef downregulates CD28	[557]
	149-155,174-179[558],158-178[559]	NBP1, ATP6V1H, V1H unit of vacuolar ATPase	133-363,402-483[559]	Subtype B	endocytic trafficking, Internalization of CD4	[558],[560],[560],
	2	ATP-binding cassette ABCA1	2225-2231[561]	HIV-1	induces downregulation, redistribution of ABCA1 to plasma membrane	[561],
	174,175,179	c-Raf1 kinase	33-38	Subtype B	Activate c-Raf1-MAP kinase pathway	[562]
	72,75[563],F195[564]	SH3 domain of Vav1 (Rho family of small GTPases)	783–843[563]	HIV-1, SIV	nef recruits Vav1 to membrane microdomains, associate with PAK2 activity	[563],[564]
	72-75	diaphanous interacting protein (DIP)	SH3 domain	HIV-1	activates p190RhoAGAP, downregulates RhoA	[565]
	72,75,78	Rack1(receptor for activated C kinase 1)	181~317	Subtype B	Rack1 acts as a Nef intracellular docking site	[566]
	69-83	mitogen-activated protein kinase (MAPK)	69-71	Subtype B	Nef decreases MAPK kinase activity in T cell	[551]
	158,177	Regulatory p85 subunit of PI3K	477-514	Subtype B	activate p21-activated kinase (PAK)	[567]
	G2,R106	apoptosis signal-regulating kinase ASK1	152-159[568]	Subtype B	Nef inhibits ASK1 pro-apoptotic signals	[568],[569]
	E62-E65[570],P72-P75,20,78,113	PACS-1 (phosphofurin acidic cluster sorting protein-1)	N188,K189,Q195	Subtype B	downregulation of MHC-I to the trans-Golgi network	[571], [572], [573],[570]
	E62-E65, P72-P75	PACS-2 (phosphofurin acidic cluster sorting protein-2)		HIV-1	Nef action and Trafficking of itinerant membrane	[574]
	E62-E65,P69,P72,P75,P78,D123[575],W13,V16,M20[576], L164,L165[577], Y143,L181,F185[57	clathrin adaptor complex 1 AP-1, μ1 subunit (AP47, AP1M1)	R303, K274E K298E K302E R303D,K374,R225, R393,K396,R211,R246,218-231[575], F172,D174,V392,L3	Subtype B,SIV	clathrin-associated sorting	[580], [575], [577], [578]

	8]		95[579]			
	E160,L164,L165[581],17-26[582],	adapter protein complex 1 AP-1, γ subunit		HIV-1	Nef interacts with γ-σ1 complex [583]	[581],[583],[577],[582], [584]
	L164,L165[556]	adapter protein complex 1 AP-1, β1 subunit (beta-adaptin)				[556]
	E160,L164,L165[584], D174-D175,E179[585]	adapter protein complex 2 AP-2, α1 subunit	K297,R340[586]	Subtype B	Nef stabilizes association of AP-1 α-σ2 complex with membranes	[584], [585]
	L164,L165[556],	adapter protein complex 2 AP-2, β1 subunit (beta-adaptin)				[556]
	E160,L164-L165[587],	clathrin adaptor complex 2 AP-2, μ1 subunit (AP50,mu2)		HIV-1,SIV,HIV-2	CD4 downregulation[587]	[587], [588]
	L164-L165	clathrin adaptor complex 3 AP-3, μ1 subunit (mu3A adaptin)		Subtype B	Nef stabilizes association of AP-1 and AP-3 with membranes	[589],
	E160,L164,L165	adapter protein complex 3 AP-3, δ subunit		HIV-1	Nef interacts with δ-σ3 complex [583],Nef stabilizes association of AP-3 with membranes	[581], [583], [584]
	R17,R19,E154,E155 [590]	Beta-COPI (component of non-clathrin-coated vesicles)		Subtype B	endosomal sorting	[590],[591],[592]
	67-69,72-73,75,77,90[549], 85,89,106,187,188,191[593],F191,F195[594]	p21 activated protein kinase PAK2		HIV-1	activate primary CD4/CD8+ T cells	[593], [595],[594]
	135-138	ALIX (ALG-2 interacting host protein, AIP-1)		Subtype B	Proliferate multivesicular body in macrophages	[596]
	D174,D175	Beclin 1		HIV-1	Prevent autophagy degradation	[597]
	66-70[598]	heat shock protein Hsp70(mortalin)		Subtype B	exosomal Nef secretion	[598],[599]
	G2-S9	ubiquitous calcium-sensing calmodulin (CaM, CAMI)		HIV-1	Nef intracellular localization and membrane targeting	[600]
	62-65,72-75	Gαi2 (α subunits of G-protein Gi)	HECT domain	Subtype B	nef recruits AIP4 E3 ligase to ubiquitinate Gαi2 for lysosomal degradation	[601]
	73-82	T Cell Receptor zeta chain(TCRξ)		Subtype B	nef-TCR form signaling extracellular complex mediates fasL upregulation	[602]
	108,112,121-123	human thioesterase II (hTE,ACOT8)		Subtype B	enhances hTE enzymatic activity	[603],[604]
	P72,P75,P106	DOCK2		HIV-1	activates Rac activity	[605]
	P72,P75,P106	ELMO1		HIV-1	activates Rac activity	[605]
	P72,P75,P106	Rac2		HIV-1	activates Rac activity	[605]
	1-57	Tumor suppressor protein p53(TP53)		Subtype B	Inhibit p53-mediated apoptosis	[606]
	16-22	p62		Subtype B	Nef phosphorylation	[607]
	186-191	β-catenin		HIV-1	Inhibit Wnt Signaling pathway	[608]
	112,121-123, 164, 165	dynamin 2 (Dyn2)	GED domain	HIV-1	Dyn2 enhances Nef activity	[609]
	6,1~35	protein kinase C- θ(PKC- theta)		HIV-1	Inhibit activated PKC binds to RACK	[610],[611]
	6	protein kinase C- δ(PKC- delta)		HIV-1	myristoylation of Nef and Pak2 activity	[610],
	14-23(alpha-helix),138-211	Embryonic ectodermal development (EED)		HIV-1	Nef recruits EED to plasma membrane	[612]
	12,13,140,141	Argonaute-2 (AGO2)		HIV-1	Nef inhibits RNA slicing activity of Ago2	[613]
	73-82	HLA-A3 heavy chains		HIV-1	T cell antigen receptor recognition	[614]
	G67,F68,P78,D123[549],W13,V16[576], 10,17-26,20,123,E62-E65,72,75,P78,	major histocompatibility complex class I MHC-I cytoplasmic domain	327-332,Y320,324,327	HIV-1	Nef downregulate MHC-I	[615], [616],[549], [575],[576], [580],
	164,165,174,175	major histocompatibility complex class II MHC-II variant (Ii) chain		Subtype B	Nef reduces mature MHC II	[617], [618]
		Naf1 (Nef-associated factor 1, VAN)	94-412	HIV-1	CD4 down-regulation on plasma membrane	[619]
	A60-E65	tumor necrosis factor receptor-associated factor TRAF2	R393,R403, F447,S454,F456	Subtype B	Activate TRAF/NF-kB and TRAF/IRF-3 pathway	[620]
	73~88	Actin				[621]
	4,7,17-22 [622], 2-27[623]	cytoplasmic membrane				[622],[623], [624]

9.3 References

1. Sauter D, Unterweger D, Vogl M, Usmani SM, Heigele A, Kluge SF, *et al.* Human tetherin exerts strong selection pressure on the HIV-1 group N Vpu protein. *PLoS Pathog* 2012,**8**:e1003093.
2. Petit SJ, Blondeau C, Towers GJ. Analysis of the human immunodeficiency virus type 1 M group Vpu domains involved in antagonizing tetherin. *J Gen Virol* 2011,**92**:2937-2948.
3. Samal AB, Ghanam RH, Fernandez TF, Monroe EB, Saad JS. NMR, biophysical, and biochemical studies reveal the minimal Calmodulin binding domain of the HIV-1 matrix protein. *J Biol Chem* 2011,**286**:33533-33543.
4. Taylor JE, Chow JY, Jeffries CM, Kwan AH, Duff AP, Hamilton WA, *et al.* Calmodulin binds a highly extended HIV-1 MA protein that refolds upon its release. *Biophys J* 2012,**103**:541-549.
5. Chow JY, Jeffries CM, Kwan AH, Guss JM, Trewhella J. Calmodulin disrupts the structure of the HIV-1 MA protein. *J Mol Biol* 2010,**400**:702-714.
6. Ghanam RH, Fernandez TF, Fledderman EL, Saad JS. Binding of calmodulin to the HIV-1 matrix protein triggers myristate exposure. *J Biol Chem* 2010,**285**:41911-41920.
7. Peytavi R, Hong SS, Gay B, d'Angeac AD, Selig L, Benichou S, *et al.* HEED, the product of the human homolog of the murine eed gene, binds to the matrix protein of HIV-1. *J Biol Chem* 1999,**274**:1635-1645.
8. Rakotobe D, Violot S, Hong SS, Gouet P, Boulanger P. Mapping of immunogenic and protein-interacting regions at the surface of the seven-bladed beta-propeller domain of the HIV-1 cellular interactor EED. *Virol J* 2008,**5**:32.
9. Giagulli C, Magiera AK, Bugatti A, Caccuri F, Marsico S, Rusnati M, *et al.* HIV-1 matrix protein p17 binds to the IL-8 receptor CXCR1 and shows IL-8-like chemokine activity on monocytes through Rho/ROCK activation. *Blood* 2012,**119**:2274-2283.
10. Caccuri F, Giagulli C, Bugatti A, Benetti A, Alessandri G, Ribatti D, *et al.* HIV-1 matrix protein p17 promotes angiogenesis via chemokine receptors CXCR1 and CXCR2. *Proc Natl Acad Sci U S A* 2012,**109**:14580-14585.
11. Bugatti A, Giagulli C, Urbinati C, Caccuri F, Chiodelli P, Oreste P, *et al.* Molecular interaction studies of HIV-1 matrix protein p17 and heparin: identification of the heparin-binding motif of p17 as a target for the development of multitarget antagonists. *J Biol Chem* 2013,**288**:1150-1161.
12. Cimarelli A, Luban J. Translation elongation factor 1-alpha interacts specifically with the human immunodeficiency virus type 1 Gag polyprotein. *J Virol* 1999,**73**:5388-5401.
13. Bristow R, Byrne J, Squirell J, Trencher H, Carter T, Rodgers B, *et al.* Human cyclophilin has a significantly higher affinity for HIV-1 recombinant p55 than p24. *J Acquir Immune Defic Syndr Hum Retrovirol* 1999,**20**:334-336.
14. Dupont S, Sharova N, DeHoratius C, Virbasius CM, Zhu X, Bukrinskaya AG, *et al.* A novel nuclear export activity in HIV-1 matrix protein required for viral replication. *Nature* 1999,**402**:681-685.
15. Haffar OK, Popov S, Dubrovsky L, Agostini I, Tang H, Pushkarsky T, *et al.* Two nuclear localization signals in the HIV-1 matrix protein regulate nuclear import of the HIV-1 pre-integration complex. *J Mol Biol* 2000,**299**:359-368.
16. Batonick M, Favre M, Boge M, Spearman P, Honing S, Thali M. Interaction of HIV-1 Gag with the clathrin-associated adaptor AP-2. *Virology* 2005,**342**:190-200.
17. Lopez-Verges S, Camus G, Blot G, Beauvoir R, Benarous R, Berlioz-Torrent C. Tail-interacting protein TIP47 is a connector between Gag and Env and is required for Env incorporation into HIV-1 virions. *Proc Natl Acad Sci U S A* 2006,**103**:14947-14952.
18. Bauby H, Lopez-Verges S, Hoeffel G, Delcroix-Genete D, Janvier K, Mammano F, *et al.* TIP47 is required for the production of infectious HIV-1 particles from primary macrophages. *Traffic* 2010,**11**:455-467.
19. Lama J, Trono D. Human immunodeficiency virus type 1 matrix protein interacts with cellular protein HO3. *J Virol* 1998,**72**:1671-1676.
20. Kaushik R, Ratner L. Role of human immunodeficiency virus type 1 matrix phosphorylation in an early postentry step of virus replication. *J Virol* 2004,**78**:2319-2326.
21. Gupta P, Singhal PK, Rajendrakumar P, Padwad Y, Tendulkar AV, Kalyanaraman VS, *et al.* Mechanism of host cell MAPK/ERK-2 incorporation into lentivirus particles: characterization

of the interaction between MAPK/ERK-2 and proline-rich-domain containing capsid region of structural protein Gag. *J Mol Biol* 2011,**410**:681-697.

22. Jacque JM, Mann A, Enslen H, Sharova N, Brichacek B, Davis RJ, *et al.* Modulation of HIV-1 infectivity by MAPK, a virion-associated kinase. *EMBO J* 1998,**17**:2607-2618.
23. Vlach J, Saad JS. Trio engagement via plasma membrane phospholipids and the myristoyl moiety governs HIV-1 matrix binding to bilayers. *Proc Natl Acad Sci U S A* 2013,**110**:3525-3530.
24. Steckbeck JD, Kuhlmann AS, Montelaro RC. C-terminal tail of human immunodeficiency virus gp41: functionally rich and structurally enigmatic. *J Gen Virol* 2013,**94**:1-19.
25. Lin CW, Engelman A. The barrier-to-autointegration factor is a component of functional human immunodeficiency virus type 1 preintegration complexes. *J Virol* 2003,**77**:5030-5036.
26. Burnette B, Yu G, Felsted RL. Phosphorylation of HIV-1 gag proteins by protein kinase C. *J Biol Chem* 1993,**268**:8698-8703.
27. Yu G, Shen FS, Sturch S, Aquino A, Glazer RI, Felsted RL. Regulation of HIV-1 gag protein subcellular targeting by protein kinase C. *J Biol Chem* 1995,**270**:4792-4796.
28. Buratti E, Tisminetzky SG, D'Agaro P, Baralle FE. A neutralizing monoclonal antibody previously mapped exclusively on human immunodeficiency virus type 1 gp41 recognizes an epitope in p17 sharing the core sequence IEEE. *J Virol* 1997,**71**:2457-2462.
29. Saad JS, Kim A, Ghanam RH, Dalton AK, Vogt VM, Wu Z, *et al.* Mutations that mimic phosphorylation of the HIV-1 matrix protein do not perturb the myristyl switch. *Protein Sci* 2007,**16**:1793-1797.
30. Saad JS, Miller J, Tai J, Kim A, Ghanam RH, Summers MF. Structural basis for targeting HIV-1 Gag proteins to the plasma membrane for virus assembly. *Proc Natl Acad Sci U S A* 2006,**103**:11364-11369.
31. Saad JS, Ablan SD, Ghanam RH, Kim A, Andrews K, Nagashima K, *et al.* Structure of the myristylated human immunodeficiency virus type 2 matrix protein and the role of phosphatidylinositol-(4,5)-bisphosphate in membrane targeting. *J Mol Biol* 2008,**382**:434-447.
32. Bennett EM, Lever AM, Allen JF. Human immunodeficiency virus type 2 Gag interacts specifically with PRP4, a serine-threonine kinase, and inhibits phosphorylation of splicing factor SF2. *J Virol* 2004,**78**:11303-11312.
33. Stauch B, Hofmann H, Perkovic M, Weisel M, Kopietz F, Cichutek K, *et al.* Model structure of APOBEC3C reveals a binding pocket modulating ribonucleic acid interaction required for encapsidation. *Proc Natl Acad Sci U S A* 2009,**106**:12079-12084.
34. Wang T, Zhang W, Tian C, Liu B, Yu Y, Ding L, *et al.* Distinct viral determinants for the packaging of human cytidine deaminases APOBEC3G and APOBEC3C. *Virology* 2008,**377**:71-79.
35. Martinez NW, Xue X, Berro RG, Kreitzer G, Resh MD. Kinesin KIF4 regulates intracellular trafficking and stability of the human immunodeficiency virus type 1 Gag polyprotein. *J Virol* 2008,**82**:9937-9950.
36. Tang Y, Winkler U, Freed EO, Torrey TA, Kim W, Li H, *et al.* Cellular motor protein KIF-4 associates with retroviral Gag. *J Virol* 1999,**73**:10508-10513.
37. Gupta K, Ott D, Hope TJ, Siliciano RF, Boeke JD. A human nuclear shuttling protein that interacts with human immunodeficiency virus type 1 matrix is packaged into virions. *J Virol* 2000,**74**:11811-11824.
38. Wilson SA, Sieiro-Vazquez C, Edwards NJ, Iourin O, Byles ED, Kotsopoulou E, *et al.* Cloning and characterization of hIF2, a human homologue of bacterial translation initiation factor 2, and its interaction with HIV-1 matrix. *Biochem J* 1999,**342 (Pt 1)**:97-103.
39. Gallay P, Stitt V, Mundy C, Oettinger M, Trono D. Role of the karyopherin pathway in human immunodeficiency virus type 1 nuclear import. *J Virol* 1996,**70**:1027-1032.
40. Gamble TR, Vajdos FF, Yoo S, Worthylake DK, Houseweart M, Sundquist WI, *et al.* Crystal structure of human cyclophilin A bound to the amino-terminal domain of HIV-1 capsid. *Cell* 1996,**87**:1285-1294.
41. Qi M, Yang R, Aiken C. Cyclophilin A-dependent restriction of human immunodeficiency virus type 1 capsid mutants for infection of nondividing cells. *J Virol* 2008,**82**:12001-12008.
42. Ylinen LM, Schaller T, Price A, Fletcher AJ, Noursadeghi M, James LC, *et al.* Cyclophilin A levels dictate infection efficiency of human immunodeficiency virus type 1 capsid escape mutants A92E and G94D. *J Virol* 2009,**83**:2044-2047.
43. Gatanaga H, Das D, Suzuki Y, Yeh DD, Hussain KA, Ghosh AK, *et al.* Altered HIV-1 Gag protein interactions with cyclophilin A (CypA) on the acquisition of H219Q and H219P substitutions in the CypA binding loop. *J Biol Chem* 2006,**281**:1241-1250.

44. Colgan J, Yuan HE, Franke EK, Luban J. Binding of the human immunodeficiency virus type 1 Gag polyprotein to cyclophilin A is mediated by the central region of capsid and requires Gag dimerization. *J Virol* 1996,**70**:4299-4310.
45. Ambrose Z, Lee K, Ndjomou J, Xu H, Oztop I, Matous J, *et al.* Human immunodeficiency virus type 1 capsid mutation N74D alters cyclophilin A dependence and impairs macrophage infection. *J Virol* 2012,**86**:4708-4714.
46. Bosco DA, Eisenmesser EZ, Clarkson MW, Wolf-Watz M, Labeikovsky W, Millet O, *et al.* Dissecting the microscopic steps of the cyclophilin A enzymatic cycle on the biological HIV-1 capsid substrate by NMR. *J Mol Biol* 2010,**403**:723-738.
47. Lin TY, Emerman M. Cyclophilin A interacts with diverse lentiviral capsids. *Retrovirology* 2006,**3**:70.
48. Mascarenhas AP, Musier-Forsyth K. The capsid protein of human immunodeficiency virus: interactions of HIV-1 capsid with host protein factors. *FEBS J* 2009,**276**:6118-6127.
49. Yoo S, Myszka DG, Yeh C, McMurray M, Hill CP, Sundquist WI. Molecular recognition in the HIV-1 capsid/cyclophilin A complex. *J Mol Biol* 1997,**269**:780-795.
50. Song C, Aiken C. Analysis of human cell heterokaryons demonstrates that target cell restriction of cyclosporine-resistant human immunodeficiency virus type 1 mutants is genetically dominant. *J Virol* 2007,**81**:11946-11956.
51. Hatziioannou T, Perez-Caballero D, Cowan S, Bieniasz PD. Cyclophilin interactions with incoming human immunodeficiency virus type 1 capsids with opposing effects on infectivity in human cells. *J Virol* 2005,**79**:176-183.
52. Price AJ, Fletcher AJ, Schaller T, Elliott T, Lee K, KewalRamani VN, *et al.* CPSF6 defines a conserved capsid interface that modulates HIV-1 replication. *PLoS Pathog* 2012,**8**:e1002896.
53. Lee K, Ambrose Z, Martin TD, Oztop I, Mulky A, Julias JG, *et al.* Flexible use of nuclear import pathways by HIV-1. *Cell Host Microbe* 2010,**7**:221-233.
54. Sayah DM, Sokolskaja E, Berthoux L, Luban J. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature* 2004,**430**:569-573.
55. Stremlau M, Perron M, Lee M, Li Y, Song B, Javanbakht H, *et al.* Specific recognition and accelerated uncoating of retroviral capsids by the TRIM5alpha restriction factor. *Proc Natl Acad Sci U S A* 2006,**103**:5514-5519.
56. Yang R, Shi J, Byeon IJ, Ahn J, Sheehan JH, Meiler J, *et al.* Second-site suppressors of HIV-1 capsid mutations: restoration of intracellular activities without correction of intrinsic capsid stability defects. *Retrovirology* 2012,**9**:30.
57. Maillard PV, Zoete V, Michielin O, Trono D. Homology-based identification of capsid determinants that protect HIV1 from human TRIM5alpha restriction. *J Biol Chem* 2011,**286**:8128-8140.
58. Li X, Song B, Xiang SH, Sodroski J. Functional interplay between the B-box 2 and the B30.2(SPRY) domains of TRIM5alpha. *Virology* 2007,**366**:234-244.
59. Li Y, Li X, Stremlau M, Lee M, Sodroski J. Removal of arginine 332 allows human TRIM5alpha to bind human immunodeficiency virus capsids and to restrict infection. *J Virol* 2006,**80**:6738-6744.
60. Diaz-Griffero F, Qin XR, Hayashi F, Kigawa T, Finzi A, Sarnak Z, *et al.* A B-box 2 surface patch important for TRIM5alpha self-association, capsid binding avidity, and retrovirus restriction. *J Virol* 2009,**83**:10737-10751.
61. Stremlau M, Owens CM, Perron MJ, Kiessling M, Autissier P, Sodroski J. The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys. *Nature* 2004,**427**:848-853.
62. De Iaco A, Luban J. Inhibition of HIV-1 infection by TNPO3 depletion is determined by capsid and detectable after viral cDNA enters the nucleus. *Retrovirology* 2011,**8**:98.
63. Zhou L, Sokolskaja E, Jolly C, James W, Cowley SA, Fassati A. Transportin 3 promotes a nuclear maturation step required for efficient HIV-1 integration. *PLoS Pathog* 2011,**7**:e1002194.
64. Valle-Casuso JC, Di Nunzio F, Yang Y, Reszka N, Lienlaf M, Arhel N, *et al.* TNPO3 is required for HIV-1 replication after nuclear import but prior to integration and binds the HIV-1 core. *J Virol* 2012,**86**:5931-5936.
65. Di Nunzio F, Danckaert A, Fricke T, Perez P, Fernandez J, Perret E, *et al.* Human nucleoporins promote HIV-1 docking at the nuclear pore, nuclear import and integration. *PLoS One* 2012,**7**:e46037.

66. Di Nunzio F, Fricke T, Miccio A, Valle-Casuso JC, Perez P, Souque P, *et al.* Nup153 and Nup98 bind the HIV-1 core and contribute to the early steps of HIV-1 replication. *Virology* 2013,**440**:8-18.
67. Matreyek KA, Yucel SS, Li X, Engelman A. Nucleoporin NUP153 Phenylalanine-Glycine Motifs Engage a Common Binding Pocket within the HIV-1 Capsid Protein to Mediate Lentiviral Infectivity. *PLoS Pathog* 2013,**9**:e1003693.
68. Matreyek KA, Engelman A. The requirement for nucleoporin NUP153 during human immunodeficiency virus type 1 infection is determined by the viral capsid. *J Virol* 2011,**85**:7818-7827.
69. Koh Y, Wu X, Ferris AL, Matreyek KA, Smith SJ, Lee K, *et al.* Differential effects of human immunodeficiency virus type 1 capsid and cellular factors nucleoporin 153 and LEDGF/p75 on the efficiency and specificity of viral DNA integration. *J Virol* 2013,**87**:648-658.
70. Bichel K, Price AJ, Schaller T, Towers GJ, Freund SM, James LC. HIV-1 capsid undergoes coupled binding and isomerization by the nuclear pore protein NUP358. *Retrovirology* 2013,**10**:81.
71. Ocwieja KE, Brady TL, Ronen K, Huegel A, Roth SL, Schaller T, *et al.* HIV integration targeting: a pathway involving Transportin-3 and the nuclear pore protein RanBP2. *PLoS Pathog* 2011,**7**:e1001313.
72. Schaller T, Ocwieja KE, Rasaiyaah J, Price AJ, Brady TL, Roth SL, *et al.* HIV-1 capsid-cyclophilin interactions determine nuclear import pathway, integration targeting and replication efficiency. *PLoS Pathog* 2011,**7**:e1002439.
73. Misumi S, Inoue M, Dochi T, Kishimoto N, Hasegawa N, Takamune N, *et al.* Uncoating of human immunodeficiency virus type 1 requires prolyl isomerase Pin1. *J Biol Chem* 2010,**285**:25185-25195.
74. Kitagawa Y, Kameoka M, Shoji-Kawata S, Iwabu Y, Mizuta H, Tokunaga K, *et al.* Inhibitory function of adapter-related protein complex 2 alpha 1 subunit in the process of nuclear translocation of human immunodeficiency virus type 1 genome. *Virology* 2008,**373**:171-180.
75. Nangola S, Urvoas A, Valerio-Lepiniec M, Khamaikawin W, Sakkhachornphop S, Hong SS, *et al.* Antiviral activity of recombinant ankyrin targeted to the capsid domain of HIV-1 Gag polyprotein. *Retrovirology* 2012,**9**:17.
76. Javanbakht H, Halwani R, Cen S, Saadatmand J, Musier-Forsyth K, Gottlinger H, *et al.* The interaction between HIV-1 Gag and human lysyl-tRNA synthetase during viral assembly. *J Biol Chem* 2003,**278**:27644-27651.
77. Henning MS, Morham SG, Goff SP, Naghavi MH. PDZD8 is a novel Gag-interacting factor that promotes retroviral infection. *J Virol* 2010,**84**:8990-8995.
78. Cooper J, Liu L, Woodruff EA, Taylor HE, Goodwin JS, D'Aquila RT, *et al.* Filamin A protein interacts with human immunodeficiency virus type 1 Gag protein and contributes to productive particle assembly. *J Biol Chem* 2011,**286**:28498-28510.
79. Abudu A, Wang X, Dang Y, Zhou T, Xiang SH, Zheng YH. Identification of molecular determinants from Moloney leukemia virus 10 homolog (MOV10) protein for virion packaging and anti-HIV-1 activity. *J Biol Chem* 2012,**287**:1220-1228.
80. Wang X, Han Y, Dang Y, Fu W, Zhou T, Ptak RG, *et al.* Moloney leukemia virus 10 (MOV10) protein inhibits retrovirus replication. *J Biol Chem* 2010,**285**:14346-14355.
81. Dussupt V, Sette P, Bello NF, Javid MP, Nagashima K, Bouamr F. Basic residues in the nucleocapsid domain of Gag are critical for late events of HIV-1 budding. *J Virol* 2011,**85**:2304-2315.
82. Zhai Q, Landesman MB, Robinson H, Sundquist WI, Hill CP. Structure of the Bro1 domain protein BROX and functional analyses of the ALIX Bro1 domain in HIV-1 budding. *PLoS One* 2011,**6**:e27466.
83. Sette P, Dussupt V, Bouamr F. Identification of the HIV-1 NC binding interface in Alix Bro1 reveals a role for RNA. *J Virol* 2012,**86**:11608-11615.
84. Dussupt V, Javid MP, Abou-Jaoude G, Jadwin JA, de La Cruz J, Nagashima K, *et al.* The nucleocapsid region of HIV-1 Gag cooperates with the PTAP and LYPXnL late domains to recruit the cellular machinery necessary for viral budding. *PLoS Pathog* 2009,**5**:e1000339.
85. Sette P, Jadwin JA, Dussupt V, Bello NF, Bouamr F. The ESCRT-associated protein Alix recruits the ubiquitin ligase Nedd4-1 to facilitate HIV-1 release through the LYPXnL L domain motif. *J Virol* 2010,**84**:8181-8192.
86. Luo K, Liu B, Xiao Z, Yu Y, Yu X, Gorelick R, *et al.* Amino-terminal region of the human immunodeficiency virus type 1 nucleocapsid is required for human APOBEC3G packaging. *J Virol* 2004,**78**:11841-11852.

87. Burnett A, Spearman P. APOBEC3G multimers are recruited to the plasma membrane for packaging into human immunodeficiency virus type 1 virus-like particles in an RNA-dependent process requiring the NC basic linker. *J Virol* 2007,**81**:5000-5013.
88. Huthoff H, Malim MH. Identification of amino acid residues in APOBEC3G required for regulation by human immunodeficiency virus type 1 Vif and Virion encapsidation. *J Virol* 2007,**81**:3807-3815.
89. Cen S, Guo F, Niu M, Saadatmand J, Deflassieux J, Kleiman L. The interaction between HIV-1 Gag and APOBEC3G. *J Biol Chem* 2004,**279**:33177-33184.
90. Song C, Sutton L, Johnson ME, D'Aquila RT, Donahue JP. Signals in APOBEC3F N-terminal and C-terminal deaminase domains each contribute to encapsidation in HIV-1 virions and are both required for HIV-1 restriction. *J Biol Chem* 2012,**287**:16965-16974.
91. Gooch BD, Cullen BR. Functional domain organization of human APOBEC3G. *Virology* 2008,**379**:118-124.
92. Zhou Y, Rong L, Lu J, Pan Q, Liang C. Insulin-like growth factor II mRNA binding protein 1 associates with Gag protein of human immunodeficiency virus type 1, and its overexpression affects virus assembly. *J Virol* 2008,**82**:5683-5692.
93. Chatel-Chaix L, Boulay K, Mouland AJ, Desgroseillers L. The host protein Staufen1 interacts with the Pr55Gag zinc fingers and regulates HIV-1 assembly via its N-terminus. *Retrovirology* 2008,**5**:41.
94. Lingappa JR, Dooher JE, Newman MA, Kiser PK, Klein KC. Basic residues in the nucleocapsid domain of Gag are required for interaction of HIV-1 gag with ABCE1 (HP68), a cellular protein important for HIV-1 capsid assembly. *J Biol Chem* 2006,**281**:3773-3784.
95. Takahashi H, Matsuda M, Kojima A, Sata T, Andoh T, Kurata T, *et al.* Human immunodeficiency virus type 1 reverse transcriptase: enhancement of activity by interaction with cellular topoisomerase I. *Proc Natl Acad Sci U S A* 1995,**92**:5694-5698.
96. Fisher RD, Chung HY, Zhai Q, Robinson H, Sundquist WI, Hill CP. Structural and biochemical studies of ALIX/AIP1 and its role in retrovirus budding. *Cell* 2007,**128**:841-852.
97. Lazert C, Chazal N, Briant L, Gerlier D, Cortay JC. Refined study of the interaction between HIV-1 p6 late domain and ALIX. *Retrovirology* 2008,**5**:39.
98. Patil A, Bhattacharya J. Natural deletion of L35Y36 in p6 gag eliminate LYPXnL/ALIX auxiliary virus release pathway in HIV-1 subtype C. *Virus Res* 2012,**170**:154-158.
99. Strack B, Calistri A, Craig S, Popova E, Gottlinger HG. AIP1/ALIX is a binding partner for HIV-1 p6 and EIAV p9 functioning in virus budding. *Cell* 2003,**114**:689-699.
100. Gurer C, Berthoux L, Luban J. Covalent modification of human immunodeficiency virus type 1 p6 by SUMO-1. *J Virol* 2005,**79**:910-917.
101. Jaber T, Bohl CR, Lewis GL, Wood C, West JT, Jr., Weldon RA, Jr. Human Ubc9 contributes to production of fully infectious human immunodeficiency virus type 1 virions. *J Virol* 2009,**83**:10448-10459.
102. Hemonnot B, Cartier C, Gay B, Rebuffat S, Bardy M, Devaux C, *et al.* The host cell MAP kinase ERK-2 regulates viral assembly and release by phosphorylating the p6gag protein of HIV-1. *J Biol Chem* 2004,**279**:32426-32434.
103. Popov S, Strack B, Sanchez-Merino V, Popova E, Rosin H, Gottlinger HG. Human immunodeficiency virus type 1 and related primate lentiviruses engage clathrin through Gag-Pol or Gag. *J Virol* 2011,**85**:3792-3801.
104. Solbak SM, Reksten TR, Roder R, Wray V, Horvli O, Raae AJ, *et al.* HIV-1 p6-Another viral interaction partner to the host cellular protein cyclophilin A. *Biochim Biophys Acta* 2012,**1824**:667-678.
105. Ott DE, Coren LV, Copeland TD, Kane BP, Johnson DG, Sowder RC, 2nd, *et al.* Ubiquitin is covalently attached to the p6Gag proteins of human immunodeficiency virus type 1 and simian immunodeficiency virus and to the p12Gag protein of Moloney murine leukemia virus. *J Virol* 1998,**72**:2962-2968.
106. Barr SD, Smiley JR, Bushman FD. The interferon response inhibits HIV particle production by induction of TRIM22. *PLoS Pathog* 2008,**4**:e1000007.
107. Ohmine S, Sakuma R, Sakuma T, Thatava T, Takeuchi H, Ikeda Y. The antiviral spectra of TRIM5alpha orthologues and human TRIM family proteins against lentiviral production. *PLoS One* 2011,**6**:e16121.
108. Shoeman RL, Hartig R, Hauses C, Traub P. Organization of focal adhesion plaques is disrupted by action of the HIV-1 protease. *Cell Biol Int* 2002,**26**:529-539.

109. Lemay J, Maidou-Peindara P, Cancio R, Ennifar E, Coadou G, Maga G, *et al.* AKAP149 binds to HIV-1 reverse transcriptase and is involved in the reverse transcription. *J Mol Biol* 2008,**383**:783-796.
110. Wang X, Ao Z, Chen L, Kobinger G, Peng J, Yao X. The cellular antiviral protein APOBEC3G interacts with HIV-1 reverse transcriptase and inhibits its function during viral replication. *J Virol* 2012,**86**:3777-3786.
111. Zhao Y, Li W, Zeng J, Liu G, Tang Y. Insights into the interactions between HIV-1 integrase and human LEDGF/p75 by molecular dynamics simulation and free energy calculation. *Proteins* 2008,**72**:635-645.
112. Cherepanov P, Ambrosio AL, Rahman S, Ellenberger T, Engelman A. Structural basis for the recognition between HIV-1 integrase and transcriptional coactivator p75. *Proc Natl Acad Sci U S A* 2005,**102**:17308-17313.
113. Rhodes DI, Peat TS, Vandegraaff N, Jeevarajah D, Newman J, Martyn J, *et al.* Crystal structures of novel allosteric peptide inhibitors of HIV integrase identify new interactions at the LEDGF binding site. *Chembiochem* 2011,**12**:2311-2315.
114. Cherepanov P, Devroe E, Silver PA, Engelman A. Identification of an evolutionarily conserved domain in human lens epithelium-derived growth factor/transcriptional co-activator p75 (LEDGF/p75) that binds HIV-1 integrase. *J Biol Chem* 2004,**279**:48883-48892.
115. Hare S, Shun MC, Gupta SS, Valkov E, Engelman A, Cherepanov P. A novel co-crystal structure affords the design of gain-of-function lentiviral integrase mutants in the presence of modified PSIP1/LEDGF/p75. *PLoS Pathog* 2009,**5**:e1000259.
116. De Houwer S, Demeulemeester J, Thys W, Taltynov O, Zmajkovicova K, Christ F, *et al.* Identification of residues in the C-terminal domain of HIV-1 integrase that mediate binding to the transportin-SR2 protein. *J Biol Chem* 2012,**287**:34059-34068.
117. Christ F, Thys W, De Rijck J, Gijsbers R, Albanese A, Arosio D, *et al.* Transportin-SR2 imports HIV into the nucleus. *Curr Biol* 2008,**18**:1192-1202.
118. Larue R, Gupta K, Wuensch C, Shkriabai N, Kessl JJ, Danhart E, *et al.* Interaction of the HIV-1 intasome with transportin 3 protein (TNPO3 or TRN-SR2). *J Biol Chem* 2012,**287**:34044-34058.
119. Levin A, Hayouka Z, Friedler A, Loyter A. Transportin 3 and importin alpha are required for effective nuclear import of HIV-1 integrase in virus-infected cells. *Nucleus* 2010,**1**:422-431.
120. Levin A, Armon-Omer A, Rosenbluh J, Melamed-Book N, Graessmann A, Waigmann E, *et al.* Inhibition of HIV-1 integrase nuclear import and replication by a peptide bearing integrase putative nuclear localization signal. *Retrovirology* 2009,**6**:112.
121. Gallay P, Hope T, Chin D, Trono D. HIV-1 infection of nondividing cells through the recognition of integrase by the importin/karyopherin pathway. *Proc Natl Acad Sci U S A* 1997,**94**:9825-9830.
122. Hearps AC, Jans DA. HIV-1 integrase is capable of targeting DNA to the nucleus via an importin alpha/beta-dependent mechanism. *Biochem J* 2006,**398**:475-484.
123. Ao Z, Danappa Jayappa K, Wang B, Zheng Y, Kung S, Rassart E, *et al.* Importin alpha3 interacts with HIV-1 integrase and contributes to HIV-1 nuclear import and replication. *J Virol* 2010,**84**:8650-8663.
124. Ao Z, Huang G, Yao H, Xu Z, Labine M, Cochrane AW, *et al.* Interaction of human immunodeficiency virus type 1 integrase with cellular nuclear import receptor importin 7 and its impact on viral replication. *J Biol Chem* 2007,**282**:13456-13467.
125. Zaitseva L, Cherepanov P, Leyens L, Wilson SJ, Rasaiyaah J, Fassati A. HIV-1 exploits importin 7 to maximize nuclear import of its DNA genome. *Retrovirology* 2009,**6**:11.
126. Mathew S, Nguyen M, Wu X, Pal A, Shah VB, Prasad VR, *et al.* INI1/hSNF5-interaction defective HIV-1 IN mutants exhibit impaired particle morphology, reverse transcription and integration in vivo. *Retrovirology* 2013,**10**:66.
127. Maroun M, Delelis O, Coadou G, Bader T, Segeral E, Mbemba G, *et al.* Inhibition of early steps of HIV-1 replication by SNF5/Ini1. *J Biol Chem* 2006,**281**:22736-22743.
128. Yung E, Sorin M, Pal A, Craig E, Morozov A, Delattre O, *et al.* Inhibition of HIV-1 virion production by a transdominant mutant of integrase interactor 1. *Nat Med* 2001,**7**:920-926.
129. Al-Mawsawi LQ, Neamati N. Blocking interactions between HIV-1 integrase and cellular cofactors: an emerging anti-retroviral strategy. *Trends Pharmacol Sci* 2007,**28**:526-535.
130. Morozov A, Yung E, Kalpana GV. Structure-function analysis of integrase interactor 1/hSNF5L1 reveals differential properties of two repeat motifs present in the highly conserved region. *Proc Natl Acad Sci U S A* 1998,**95**:1120-1125.

131. Das S, Cano J, Kalpana GV. Multimerization and DNA binding properties of INI1/hSNF5 and its functional significance. *J Biol Chem* 2009,**284**:19903-19914.
132. Cereseto A, Manganaro L, Gutierrez MI, Terreni M, Fittipaldi A, Lusic M, *et al.* Acetylation of HIV-1 integrase by p300 regulates viral integration. *EMBO J* 2005,**24**:3070-3081.
133. Topper M, Luo Y, Zhadina M, Mohammed K, Smith L, Muesing MA. Posttranslational acetylation of the human immunodeficiency virus type 1 integrase carboxyl-terminal domain is dispensable for viral replication. *J Virol* 2007,**81**:3012-3017.
134. Nishitsuji H, Hayashi T, Takahashi T, Miyano M, Kannagi M, Masuda T. Augmentation of reverse transcription by integrase through an interaction with host factor, SIP1/Gemin2 Is critical for HIV-1 infection. *PLoS One* 2009,**4**:e7825.
135. Hamamoto S, Nishitsuji H, Amagasa T, Kannagi M, Masuda T. Identification of a novel human immunodeficiency virus type 1 integrase interactor, Gemin2, that facilitates efficient viral cDNA synthesis in vivo. *J Virol* 2006,**80**:5670-5677.
136. Mousnier A, Kubat N, Massias-Simon A, Segeral E, Rain JC, Benarous R, *et al.* von Hippel Lindau binding protein 1-mediated degradation of integrase affects HIV-1 gene expression at a postintegration step. *Proc Natl Acad Sci U S A* 2007,**104**:13615-13620.
137. Yamamoto SP, Okawa K, Nakano T, Sano K, Ogawa K, Masuda T, *et al.* Huwe1, a novel cellular interactor of Gag-Pol through integrase binding, negatively influences HIV-1 infectivity. *Microbes Infect* 2011,**13**:339-349.
138. Violot S, Hong SS, Rakotobe D, Petit C, Gay B, Moreau K, *et al.* The human polycomb group EED protein interacts with the integrase of human immunodeficiency virus type 1. *J Virol* 2003,**77**:12507-12522.
139. Parissi V, Calmels C, De Soultrait VR, Caumont A, Fournier M, Chaignepain S, *et al.* Functional interactions of human immunodeficiency virus type 1 integrase with human and yeast HSP60. *J Virol* 2001,**75**:11344-11353.
140. Priet S, Navarro JM, Gros N, Querat G, Sire J. Functional role of HIV-1 virion-associated uracil DNA glycosylase 2 in the correction of G:U mispairs to G:C pairs. *J Biol Chem* 2003,**278**:4566-4571.
141. Willetts KE, Rey F, Agostini I, Navarro JM, Baudat Y, Vigne R, *et al.* DNA repair enzyme uracil DNA glycosylase is specifically incorporated into human immunodeficiency virus type 1 viral particles through a Vpr-independent mechanism. *J Virol* 1999,**73**:1682-1688.
142. Allouch A, Di Primio C, Alpi E, Lusic M, Arosio D, Giacca M, *et al.* The TRIM family protein KAP1 inhibits HIV-1 integration. *Cell Host Microbe* 2011,**9**:484-495.
143. Terreni M, Valentini P, Liverani V, Gutierrez MI, Di Primio C, Di Fenza A, *et al.* GCN5-dependent acetylation of HIV-1 integrase enhances viral integration. *Retrovirology* 2010,**7**:18.
144. Zheng Y, Ao Z, Wang B, Jayappa KD, Yao X. Host protein Ku70 binds and protects HIV-1 integrase from proteasomal degradation and is required for HIV replication. *J Biol Chem* 2011,**286**:17722-17735.
145. Zhang F, Zang T, Wilson SJ, Johnson MC, Bieniasz PD. Clathrin facilitates the morphogenesis of retrovirus particles. *PLoS Pathog* 2011,**7**:e1002119.
146. Manganaro L, Lusic M, Gutierrez MI, Cereseto A, Del Sal G, Giacca M. Concerted action of cellular JNK and Pin1 restricts HIV-1 genome integration to activated CD4+ T lymphocytes. *Nat Med* 2010,**16**:329-333.
147. Zamborlini A, Coiffic A, Beauclair G, Delelis O, Paris J, Koh Y, *et al.* Impairment of human immunodeficiency virus type-1 integrase SUMOylation correlates with an early replication defect. *J Biol Chem* 2011,**286**:21013-21022.
148. Li WJ, Huang L, Zhang JQ, Xu GL, Tian L, Xue JL, *et al.* The 156KELK159 tetrapeptide of HIV-1 integrase is critical for lentiviral gene integration. *Mol Biol Rep* 2012,**39**:343-349.
149. Huang L, Xu GL, Zhang JQ, Tian L, Xue JL, Chen JZ, *et al.* Daxx interacts with HIV-1 integrase and inhibits lentiviral gene expression. *Biochem Biophys Res Commun* 2008,**373**:241-245.
150. Wielens J, Headey SJ, Jeevarajah D, Rhodes DI, Deadman J, Chalmers DK, *et al.* Crystal structure of the HIV-1 integrase core domain in complex with sucrose reveals details of an allosteric inhibitory binding site. *FEBS Lett* 2010,**584**:1455-1462.
151. Luo K, Wang T, Liu B, Tian C, Xiao Z, Kappes J, *et al.* Cytidine deaminases APOBEC3G and APOBEC3F interact with human immunodeficiency virus type 1 integrase and inhibit proviral DNA formation. *J Virol* 2007,**81**:7238-7248.
152. Vandegraaff N, Devroe E, Turlure F, Silver PA, Engelman A. Biochemical and genetic analyses of integrase-interacting proteins lens epithelium-derived growth factor (LEDGF)/p75

and hepatoma-derived growth factor related protein 2 (HRP2) in preintegration complex function and HIV-1 replication. *Virology* 2006,**346**:415-426.

153. Vanegas M, Llano M, Delgado S, Thompson D, Peretz M, Poeschla E. Identification of the LEDGF/p75 HIV-1 integrase-interaction domain and NLS reveals NLS-independent chromatin tethering. *J Cell Sci* 2005,**118**:1733-1743.
154. Mulder LC, Chakrabarti LA, Muesing MA. Interaction of HIV-1 integrase with DNA repair protein hRad18. *J Biol Chem* 2002,**277**:27489-27493.
155. Zhang JQ, Wang JJ, Li WJ, Huang L, Tian L, Xue JL, *et al.* Cellular protein TTRAP interacts with HIV-1 integrase to facilitate viral integration. *Biochem Biophys Res Commun* 2009,**387**:256-260.
156. Woodward CL, Prakobwanakit S, Mosessian S, Chow SA. Integrase interacts with nucleoporin NUP153 to mediate the nuclear import of human immunodeficiency virus type 1. *J Virol* 2009,**83**:6522-6533.
157. Ao Z, Jayappa KD, Wang B, Zheng Y, Wang X, Peng J, *et al.* Contribution of host nucleoporin 62 in HIV-1 integrase chromatin association and viral DNA integration. *J Biol Chem* 2012,**287**:10544-10555.
158. Kitamura S, Ode H, Iwatani Y. Structural Features of Antiviral APOBEC3 Proteins are Linked to Their Functional Activities. *Front Microbiol* 2011,**2**:258.
159. Chen G, He Z, Wang T, Xu R, Yu XF. A patch of positively charged amino acids surrounding the human immunodeficiency virus type 1 Vif SLVx4Yx9Y motif influences its interaction with APOBEC3G. *J Virol* 2009,**83**:8674-8682.
160. Albin JS, Harris RS. Interactions of host APOBEC3 restriction factors with HIV-1 in vivo: implications for therapeutics. *Expert Rev Mol Med* 2010,**12**:e4.
161. Kitamura S, Ode H, Nakashima M, Imahashi M, Naganawa Y, Kurosawa T, *et al.* The APOBEC3C crystal structure and the interface for HIV-1 Vif binding. *Nat Struct Mol Biol* 2012,**19**:1005-1010.
162. Albin JS, Hache G, Hultquist JF, Brown WL, Harris RS. Long-term restriction by APOBEC3F selects human immunodeficiency virus type 1 variants with restored Vif function. *J Virol* 2010,**84**:10209-10219.
163. Simon V, Zennou V, Murray D, Huang Y, Ho DD, Bieniasz PD. Natural variation in Vif: differential impact on APOBEC3G/3F and a potential role in HIV-1 diversification. *PLoS Pathog* 2005,**1**:e6.
164. Dang Y, Davis RW, York IA, Zheng YH. Identification of 81LGxGxxIxW89 and 171EDRW174 domains from human immunodeficiency virus type 1 Vif that regulate APOBEC3G and APOBEC3F neutralizing activity. *J Virol* 2010,**84**:5741-5750.
165. Bohn MF, Shandilya SM, Albin JS, Kouno T, Anderson BD, McDougle RM, *et al.* Crystal structure of the DNA cytosine deaminase APOBEC3F: the catalytically active and HIV-1 Vif-binding domain. *Structure* 2013,**21**:1042-1050.
166. Smith JL, Pathak VK. Identification of specific determinants of human APOBEC3F, APOBEC3C, and APOBEC3DE and African green monkey APOBEC3F that interact with HIV-1 Vif. *J Virol* 2010,**84**:12599-12608.
167. Dang Y, Wang X, York IA, Zheng YH. Identification of a critical T(Q/D/E)x5ADx2(I/L) motif from primate lentivirus Vif proteins that regulate APOBEC3G and APOBEC3F neutralizing activity. *J Virol* 2010,**84**:8561-8570.
168. Russell RA, Pathak VK. Identification of two distinct human immunodeficiency virus type 1 Vif determinants critical for interactions with human APOBEC3G and APOBEC3F. *J Virol* 2007,**81**:8201-8210.
169. Iwatani Y, Chan DS, Liu L, Yoshii H, Shibata J, Yamamoto N, *et al.* HIV-1 Vif-mediated ubiquitination/degradation of APOBEC3G involves four critical lysine residues in its C-terminal domain. *Proc Natl Acad Sci U S A* 2009,**106**:19539-19544.
170. Donahue JP, Vetter ML, Mukhtar NA, D'Aquila RT. The HIV-1 Vif PPLP motif is necessary for human APOBEC3G binding and degradation. *Virology* 2008,**377**:49-53.
171. Xu H, Svarovskaia ES, Barr R, Zhang Y, Khan MA, Strebel K, *et al.* A single amino acid substitution in human APOBEC3G antiretroviral enzyme confers resistance to HIV-1 virion infectivity factor-induced depletion. *Proc Natl Acad Sci U S A* 2004,**101**:5652-5657.
172. Yu X, Yu Y, Liu B, Luo K, Kong W, Mao P, *et al.* Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex. *Science* 2003,**302**:1056-1060.
173. Zhen A, Wang T, Zhao K, Xiong Y, Yu XF. A single amino acid difference in human APOBEC3H variants determines HIV-1 Vif sensitivity. *J Virol* 2010,**84**:1902-1911.

174. Mehle A, Goncalves J, Santa-Marta M, McPike M, Gabuzda D. Phosphorylation of a novel SOCS-box regulates assembly of the HIV-1 Vif-Cul5 complex that promotes APOBEC3G degradation. *Genes Dev* 2004,**18**:2861-2866.
175. Mehle A, Thomas ER, Rajendran KS, Gabuzda D. A zinc-binding region in Vif binds Cul5 and determines cullin selection. *J Biol Chem* 2006,**281**:17259-17265.
176. Stanley BJ, Ehrlich ES, Short L, Yu Y, Xiao Z, Yu XF, *et al.* Structural insight into the human immunodeficiency virus Vif SOCS box and its role in human E3 ubiquitin ligase assembly. *J Virol* 2008,**82**:8656-8663.
177. Bergeron JR, Huthoff H, Veselkov DA, Beavil RL, Simpson PJ, Matthews SJ, *et al.* The SOCS-box of HIV-1 Vif interacts with ElonginBC by induced-folding to recruit its Cul5-containing ubiquitin ligase complex. *PLoS Pathog* 2010,**6**:e1000925.
178. Izumi T, Takaori-Kondo A, Shirakawa K, Higashitsuji H, Itoh K, Io K, *et al.* MDM2 is a novel E3 ligase for HIV-1 Vif. *Retrovirology* 2009,**6**:1.
179. Dussart S, Courcoul M, Bessou G, Douaisi M, Duverger Y, Vigne R, *et al.* The Vif protein of human immunodeficiency virus type 1 is posttranslationally modified by ubiquitin. *Biochem Biophys Res Commun* 2004,**315**:66-72.
180. Hassaine G, Courcoul M, Bessou G, Barthalay Y, Picard C, Olive D, *et al.* The tyrosine kinase Hck is an inhibitor of HIV-1 replication counteracted by the viral vif protein. *J Biol Chem* 2001,**276**:16885-16893.
181. Marcsisin SR, Narute PS, Emert-Sedlak LA, Kloczewiak M, Smithgall TE, Engen JR. On the solution conformation and dynamics of the HIV-1 viral infectivity factor. *J Mol Biol* 2011,**410**:1008-1022.
182. Wang X, Wang X, Zhang H, Lv M, Zuo T, Wu H, *et al.* Interactions between HIV-1 Vif and human ElonginB-ElonginC are important for CBF-beta binding to Vif. *Retrovirology* 2013,**10**:94.
183. Yu Y, Xiao Z, Ehrlich ES, Yu X, Yu XF. Selective assembly of HIV-1 Vif-Cul5-ElonginB-ElonginC E3 ubiquitin ligase complex through a novel SOCS box and upstream cysteines. *Genes Dev* 2004,**18**:2867-2872.
184. Batisse J, Guerrero S, Bernacchi S, Sleiman D, Gabus C, Darlix JL, *et al.* The role of Vif oligomerization and RNA chaperone activity in HIV-1 replication. *Virus Res* 2012,**169**:361-376.
185. Yang X, Gabuzda D. Mitogen-activated protein kinase phosphorylates and regulates the HIV-1 Vif protein. *J Biol Chem* 1998,**273**:29879-29887.
186. Izumi T, Io K, Matsui M, Shirakawa K, Shinohara M, Nagai Y, *et al.* HIV-1 viral infectivity factor interacts with TP53 to induce G2 cell cycle arrest and positively regulate viral replication. *Proc Natl Acad Sci U S A* 2010,**107**:20798-20803.
187. Zhang W, Du J, Evans SL, Yu Y, Yu XF. T-cell differentiation factor CBF-beta regulates HIV-1 Vif-mediated evasion of host restriction. *Nature* 2012,**481**:376-379.
188. Hultquist JF, McDougle RM, Anderson BD, Harris RS. HIV type 1 viral infectivity factor and the RUNX transcription factors interact with core binding factor beta on genetically distinct surfaces. *AIDS Res Hum Retroviruses* 2012,**28**:1543-1551.
189. Madani N, Millette R, Platt EJ, Marin M, Kozak SL, Bloch DB, *et al.* Implication of the lymphocyte-specific nuclear body protein Sp140 in an innate response to human immunodeficiency virus type 1. *J Virol* 2002,**76**:11133-11138.
190. Levin A, Rosenbluh J, Hayouka Z, Friedler A, Loyter A. Integration of HIV-1 DNA is regulated by interplay between viral rev and cellular LEDGF/p75 proteins. *Mol Med* 2010,**16**:34-44.
191. Bogerd HP, Fridell RA, Madore S, Cullen BR. Identification of a novel cellular cofactor for the Rev/Rex class of retroviral regulatory proteins. *Cell* 1995,**82**:485-494.
192. Venkatesh LK, Gettemeier T, Chinnadurai G. A nuclear kinesin-like protein interacts with and stimulates the activity of the leucine-rich nuclear export signal of the human immunodeficiency virus type 1 rev protein. *J Virol* 2003,**77**:7236-7243.
193. Stutz F, Neville M, Rosbash M. Identification of a novel nuclear pore-associated protein as a functional target of the HIV-1 Rev protein in yeast. *Cell* 1995,**82**:495-506.
194. Sanchez-Velar N, Udofia EB, Yu Z, Zapp ML. hRIP, a cellular cofactor for Rev function, promotes release of HIV RNAs from the perinuclear region. *Genes Dev* 2004,**18**:23-34.
195. Stutz F, Izaurralde E, Mattaj IW, Rosbash M. A role for nucleoporin FG repeat domains in export of human immunodeficiency virus type 1 Rev protein and RNA from the nucleus. *Mol Cell Biol* 1996,**16**:7144-7150.

196. Ma YL, Peng JY, Zhang P, Huang L, Liu WJ, Shen TY, *et al.* Heterogeneous nuclear ribonucleoprotein A1 is identified as a potential biomarker for colorectal cancer based on differential proteomics technology. *J Proteome Res* 2009,**8**:4525-4535.
197. Li J, Liu Y, Park IW, He JJ. Expression of exogenous Sam68, the 68-kilodalton SRC-associated protein in mitosis, is able to alleviate impaired Rev function in astrocytes. *J Virol* 2002,**76**:4526-4535.
198. Zhang J, Liu Y, Henao J, Rugeles MT, Li J, Chen T, *et al.* Requirement of an additional Sam68 domain for inhibition of human immunodeficiency virus type 1 replication by Sam68 dominant negative mutants lacking the nuclear localization signal. *Gene* 2005,**363**:67-76.
199. Kaminski R, Darbinian N, Sawaya BE, Slonina D, Amini S, Johnson EM, *et al.* Puralpha as a cellular co-factor of Rev/RRE-mediated expression of HIV-1 intron-containing mRNA. *J Cell Biochem* 2008,**103**:1231-1245.
200. Neville M, Stutz F, Lee L, Davis LI, Rosbash M. The importin-beta family member Crm1p bridges the interaction between Rev and the nuclear pore complex during nuclear export. *Curr Biol* 1997,**7**:767-775.
201. Edgcomb SP, Carmel AB, Naji S, Ambrus-Aikelin G, Reyes JR, Saphire AC, *et al.* DDX1 is an RNA-dependent ATPase involved in HIV-1 Rev function and virus replication. *J Mol Biol* 2012,**415**:61-74.
202. Fang J, Kubota S, Yang B, Zhou N, Zhang H, Godbout R, *et al.* A DEAD box protein facilitates HIV-1 replication as a cellular co-factor of Rev. *Virology* 2004,**330**:471-480.
203. Robertson-Anderson RM, Wang J, Edgcomb SP, Carmel AB, Williamson JR, Millar DP. Single-molecule studies reveal that DEAD box protein DDX1 promotes oligomerization of HIV-1 Rev on the Rev response element. *J Mol Biol* 2011,**410**:959-971.
204. Yasuda-Inoue M, Kuroki M, Ariumi Y. Distinct DDX DEAD-box RNA helicases cooperate to modulate the HIV-1 Rev function. *Biochem Biophys Res Commun* 2013,**434**:803-808.
205. Naji S, Ambrus G, Cimermancic P, Reyes JR, Johnson JR, Filbrandt R, *et al.* Host cell interactome of HIV-1 Rev includes RNA helicases involved in multiple facets of virus production. *Mol Cell Proteomics* 2012,**11**:M111 015313.
206. Kiss A, Li L, Gettemeier T, Venkatesh LK. Functional analysis of the interaction of the human immunodeficiency virus type 1 Rev nuclear export signal with its cofactors. *Virology* 2003,**314**:591-600.
207. Zolotukhin AS, Felber BK. Nucleoporins nup98 and nup214 participate in nuclear export of human immunodeficiency virus type 1 Rev. *J Virol* 1999,**73**:120-127.
208. !!! INVALID CITATION !!!
209. Truant R, Cullen BR. The arginine-rich domains present in human immunodeficiency virus type 1 Tat and Rev function as direct importin beta-dependent nuclear localization signals. *Mol Cell Biol* 1999,**19**:1210-1217.
210. Arnold M, Nath A, Hauber J, Kehlenbach RH. Multiple importins function as nuclear transport receptors for the Rev protein of human immunodeficiency virus type 1. *J Biol Chem* 2006,**281**:20883-20890.
211. Henderson BR, Percipalle P. Interactions between HIV Rev and nuclear import and export factors: the Rev nuclear localisation signal mediates specific binding to human importin-beta. *J Mol Biol* 1997,**274**:693-707.
212. Watts NR, Sackett DL, Ward RD, Miller MW, Wingfield PT, Stahl SS, *et al.* HIV-1 rev depolymerizes microtubules to form stable bilayered rings. *J Cell Biol* 2000,**150**:349-360.
213. Gu L, Tsuji T, Jarboui MA, Yeo GP, Sheehy N, Hall WW, *et al.* Intermolecular masking of the HIV-1 Rev NLS by the cellular protein HIC: novel insights into the regulation of Rev nuclear import. *Retrovirology* 2011,**8**:17.
214. Cochrane A, Murley LL, Gao M, Wong R, Clayton K, Brufatto N, *et al.* Stable complex formation between HIV Rev and the nucleosome assembly protein, NAP1, affects Rev function. *Virology* 2009,**388**:103-111.
215. Szebeni A, Hingorani K, Negi S, Olson MO. Role of protein kinase CK2 phosphorylation in the molecular chaperone activity of nucleolar protein b23. *J Biol Chem* 2003,**278**:9107-9115.
216. Li YP. Protein B23 is an important human factor for the nucleolar localization of the human immunodeficiency virus protein Tat. *J Virol* 1997,**71**:4098-4102.
217. Fankhauser C, Izaurralde E, Adachi Y, Wingfield P, Laemmli UK. Specific complex of human immunodeficiency virus type 1 rev and nucleolar B23 proteins: dissociation by the Rev response element. *Mol Cell Biol* 1991,**11**:2567-2575.

218. Tange TO, Jensen TH, Kjems J. In vitro interaction between human immunodeficiency virus type 1 Rev protein and splicing factor ASF/SF2-associated protein, p32. *J Biol Chem* 1996,**271**:10066-10072.
219. Meggio F, D'Agostino DM, Ciminale V, Chieco-Bianchi L, Pinna LA. Phosphorylation of HIV-1 Rev protein: implication of protein kinase CK2 and pro-directed kinases. *Biochem Biophys Res Commun* 1996,**226**:547-554.
220. Meggio F, Marin O, Boschetti M, Sarno S, Pinna LA. HIV-1 Rev transactivator: a beta-subunit directed substrate and effector of protein kinase CK2. *Mol Cell Biochem* 2001,**227**:145-151.
221. Kubota S, Adachi Y, Copeland TD, Oroszlan S. Binding of human prothymosin alpha to the leucine-motif/activation domains of HTLV-I Rex and HIV-1 Rev. *Eur J Biochem* 1995,**233**:48-54.
222. Farjot G, Sergeant A, Mikaelian I. A new nucleoporin-like protein interacts with both HIV-1 Rev nuclear export signal and CRM-1. *J Biol Chem* 1999,**274**:17309-17317.
223. Ruhl M, Himmelspach M, Bahr GM, Hammerschmid F, Jaksche H, Wolff B, *et al.* Eukaryotic initiation factor 5A is a cellular target of the human immunodeficiency virus type 1 Rev activation domain mediating trans-activation. *J Cell Biol* 1993,**123**:1309-1320.
224. Bevec D, Jaksche H, Oft M, Wohl T, Himmelspach M, Pacher A, *et al.* Inhibition of HIV-1 replication in lymphocytes by mutants of the Rev cofactor eIF-5A. *Science* 1996,**271**:1858-1860.
225. Hofmann W, Reichart B, Ewald A, Muller E, Schmitt I, Stauber RH, *et al.* Cofactor requirements for nuclear export of Rev response element (RRE)- and constitutive transport element (CTE)-containing retroviral RNAs. An unexpected role for actin. *J Cell Biol* 2001,**152**:895-910.
226. Modem S, Reddy TR. An anti-apoptotic protein, Hax-1, inhibits the HIV-1 rev function by altering its sub-cellular localization. *J Cell Physiol* 2008,**214**:14-19.
227. Zhou Y, Rong L, Zhang J, Aloysius C, Pan Q, Liang C. Insulin-like growth factor II mRNA binding protein 1 modulates Rev-dependent human immunodeficiency virus type 1 RNA expression. *Virology* 2009,**393**:210-220.
228. Zhou X, Luo J, Mills L, Wu S, Pan T, Geng G, *et al.* DDX5 facilitates HIV-1 replication as a cellular co-factor of Rev. *PLoS One* 2013,**8**:e65040.
229. Vigan R, Neil SJ. Determinants of tetherin antagonism in the transmembrane domain of the human immunodeficiency virus type 1 Vpu protein. *J Virol* 2010,**84**:12958-12970.
230. Malim MH, Bieniasz PD. HIV Restriction Factors and Mechanisms of Evasion. *Cold Spring Harb Perspect Med* 2012,**2**:a006940.
231. Lim ES, Malik HS, Emerman M. Ancient adaptive evolution of tetherin shaped the functions of Vpu and Nef in human immunodeficiency virus and primate lentiviruses. *J Virol* 2010,**84**:7124-7134.
232. Kobayashi T, Ode H, Yoshida T, Sato K, Gee P, Yamamoto SP, *et al.* Identification of amino acids in the human tetherin transmembrane domain responsible for HIV-1 Vpu interaction and susceptibility. *J Virol* 2011,**85**:932-945.
233. Dutta S, Tan YJ. Structural and functional characterization of human SGT and its interaction with Vpu of the human immunodeficiency virus type 1. *Biochemistry* 2008,**47**:10123-10131.
234. Callahan MA, Handley MA, Lee YH, Talbot KJ, Harper JW, Panganiban AT. Functional interaction of human immunodeficiency virus type 1 Vpu and Gag with a novel member of the tetratricopeptide repeat protein family. *J Virol* 1998,**72**:5189-5197.
235. Magadan JG, Bonifacino JS. Transmembrane domain determinants of CD4 Downregulation by HIV-1 Vpu. *J Virol* 2012,**86**:757-772.
236. Margottin F, Benichou S, Durand H, Richard V, Liu LX, Gomas E, *et al.* Interaction between the cytoplasmic domains of HIV-1 Vpu and CD4: role of Vpu residues involved in CD4 interaction and in vitro CD4 degradation. *Virology* 1996,**223**:381-386.
237. Schubert U, Strebel K. Differential activities of the human immunodeficiency virus type 1-encoded Vpu protein are regulated by phosphorylation and occur in different cellular compartments. *J Virol* 1994,**68**:2260-2271.
238. Henklein P, Schubert U, Kunert O, Klabunde S, Wray V, Kloppel KD, *et al.* Synthesis and characterization of the hydrophilic C-terminal domain of the human immunodeficiency virus type 1-encoded virus protein U (Vpu). *Pept Res* 1993,**6**:79-87.
239. Coadou G, Gharbi-Benarous J, Megy S, Bertho G, Evrard-Todeschi N, Segeral E, *et al.* NMR studies of the phosphorylation motif of the HIV-1 protein Vpu bound to the F-box protein beta-TrCP. *Biochemistry* 2003,**42**:14741-14751.

240. Estrabaud E, Le Rouzic E, Lopez-Verges S, Morel M, Belaidouni N, Benarous R, *et al.* Regulated degradation of the HIV-1 Vpu protein through a betaTrCP-independent pathway limits the release of viral particles. *PLoS Pathog* 2007,**3**:e104.
241. Margottin F, Bour SP, Durand H, Selig L, Benichou S, Richard V, *et al.* A novel human WD protein, h-beta TrCp, that interacts with HIV-1 Vpu connects CD4 to the ER degradation pathway through an F-box motif. *Mol Cell* 1998,**1**:565-574.
242. Doehle BP, Chang K, Rustagi A, McNevin J, McElrath MJ, Gale M, Jr. Vpu mediates depletion of interferon regulatory factor 3 during HIV infection by a lysosome-dependent mechanism. *J Virol* 2012,**86**:8367-8374.
243. Hsu K, Seharaseyon J, Dong P, Bour S, Marban E. Mutual functional destruction of HIV-1 Vpu and host TASK-1 channel. *Mol Cell* 2004,**14**:259-267.
244. Emeagwali N, Hildreth JE. Human immunodeficiency virus type 1 Vpu and cellular TASK proteins suppress transcription of unintegrated HIV-1 DNA. *Virol J* 2012,**9**:277.
245. Le Rouzic E, Mousnier A, Rustum C, Stutz F, Hallberg E, Dargemont C, *et al.* Docking of HIV-1 Vpr to the nuclear envelope is mediated by the interaction with the nucleoporin hCG1. *J Biol Chem* 2002,**277**:45091-45098.
246. Jacquot G, Le Rouzic E, David A, Mazzolini J, Bouchet J, Bouaziz S, *et al.* Localization of HIV-1 Vpr to the nuclear envelope: impact on Vpr functions and virus replication in macrophages. *Retrovirology* 2007,**4**:84.
247. Kamata M, Nitahara-Kasahara Y, Miyamoto Y, Yoneda Y, Aida Y. Importin-alpha promotes passage through the nuclear pore complex of human immunodeficiency virus type 1 Vpr. *J Virol* 2005,**79**:3557-3564.
248. Takeda E, Murakami T, Matsuda G, Murakami H, Zako T, Maeda M, *et al.* Nuclear exportin receptor CAS regulates the NPI-1-mediated nuclear import of HIV-1 Vpr. *PLoS One* 2011,**6**:e27815.
249. Popov S, Rexach M, Zybarth G, Reiling N, Lee MA, Ratner L, *et al.* Viral protein R regulates nuclear import of the HIV-1 pre-integration complex. *EMBO J* 1998,**17**:909-917.
250. Nitahara-Kasahara Y, Kamata M, Yamamoto T, Zhang X, Miyamoto Y, Muneta K, *et al.* Novel nuclear import of Vpr promoted by importin alpha is crucial for human immunodeficiency virus type 1 replication in macrophages. *J Virol* 2007,**81**:5284-5293.
251. Caly L, Saksena NK, Piller SC, Jans DA. Impaired nuclear import and viral incorporation of Vpr derived from a HIV long-term non-progressor. *Retrovirology* 2008,**5**:67.
252. Jenkins Y, McEntee M, Weis K, Greene WC. Characterization of HIV-1 vpr nuclear import: analysis of signals and pathways. *J Cell Biol* 1998,**143**:875-885.
253. Kino T, Gragerov A, Slobodskaya O, Tsopanomichalou M, Chrousos GP, Pavlakis GN. Human immunodeficiency virus type 1 (HIV-1) accessory protein Vpr induces transcription of the HIV-1 and glucocorticoid-responsive promoters by binding directly to p300/CBP coactivators. *J Virol* 2002,**76**:9724-9734.
254. Selig L, Benichou S, Rogel ME, Wu LI, Vodicka MA, Sire J, *et al.* Uracil DNA glycosylase specifically interacts with Vpr of both human immunodeficiency virus type 1 and simian immunodeficiency virus of sooty mangabeys, but binding does not correlate with cell cycle arrest. *J Virol* 1997,**71**:4842-4846.
255. BouHamdan M, Xue Y, Baudat Y, Hu B, Sire J, Pomerantz RJ, *et al.* Diversity of HIV-1 Vpr interactions involves usage of the WXXF motif of host cell proteins. *J Biol Chem* 1998,**273**:8009-8016.
256. Mansky LM, Preveral S, Selig L, Benarous R, Benichou S. The interaction of vpr with uracil DNA glycosylase modulates the human immunodeficiency virus type 1 In vivo mutation rate. *J Virol* 2000,**74**:7039-7047.
257. Tan L, Ehrlich E, Yu XF. DDB1 and Cul4A are required for human immunodeficiency virus type 1 Vpr-induced G2 arrest. *J Virol* 2007,**81**:10822-10830.
258. Zhao LJ, Mukherjee S, Narayan O. Biochemical mechanism of HIV-I Vpr function. Specific interaction with a cellular protein. *J Biol Chem* 1994,**269**:15577-15582.
259. Zhang S, Feng Y, Narayan O, Zhao LJ. Cytoplasmic retention of HIV-1 regulatory protein Vpr by protein-protein interaction with a novel human cytoplasmic protein VprBP. *Gene* 2001,**263**:131-140.
260. Le Rouzic E, Belaidouni N, Estrabaud E, Morel M, Rain JC, Transy C, *et al.* HIV1 Vpr arrests the cell cycle by recruiting DCAF1/VprBP, a receptor of the Cul4-DDB1 ubiquitin ligase. *Cell Cycle* 2007,**6**:182-188.

261. Yedavalli VS, Shih HM, Chiang YP, Lu CY, Chang LY, Chen MY, *et al.* Human immunodeficiency virus type 1 Vpr interacts with antiapoptotic mitochondrial protein HAX-1. *J Virol* 2005,**79**:13735-13746.
262. Jacotot E, Ferri KF, El Hamel C, Brenner C, Druillennec S, Hoebeke J, *et al.* Control of mitochondrial membrane permeabilization by adenine nucleotide translocator interacting with HIV-1 viral protein rR and Bcl-2. *J Exp Med* 2001,**193**:509-519.
263. Jacotot E, Ravagnan L, Loeffler M, Ferri KF, Vieira HL, Zamzami N, *et al.* The HIV-1 viral protein R induces apoptosis via a direct effect on the mitochondrial permeability transition pore. *J Exp Med* 2000,**191**:33-46.
264. Kamata M, Watanabe N, Nagaoka Y, Chen IS. Human immunodeficiency virus type 1 Vpr binds to the N lobe of the Wee1 kinase domain and enhances kinase activity for CDC2. *J Virol* 2008,**82**:5672-5682.
265. Mansky LM, Preveral S, Le Rouzic E, Bernard LC, Selig L, Depienne C, *et al.* Interaction of human immunodeficiency virus type 1 Vpr with the HHR23A DNA repair protein does not correlate with multiple biological functions of Vpr. *Virology* 2001,**282**:176-185.
266. Withers-Ward ES, Jowett JB, Stewart SA, Xie YM, Garfinkel A, Shibagaki Y, *et al.* Human immunodeficiency virus type 1 Vpr interacts with HHR23A, a cellular protein implicated in nucleotide excision DNA repair. *J Virol* 1997,**71**:9732-9742.
267. Withers-Ward ES, Mueller TD, Chen IS, Feigon J. Biochemical and structural analysis of the interaction between the UBA(2) domain of the DNA repair protein HHR23A and HIV-1 Vpr. *Biochemistry* 2000,**39**:14103-14112.
268. Kino T, Gragerov A, Kopp JB, Stauber RH, Pavlakis GN, Chrousos GP. The HIV-1 virion-associated protein vpr is a coactivator of the human glucocorticoid receptor. *J Exp Med* 1999,**189**:51-62.
269. Agostini I, Navarro JM, Rey F, Bouhamdan M, Spire B, Vigne R, *et al.* The human immunodeficiency virus type 1 Vpr transactivator: cooperation with promoter-bound activator domains and binding to TFIIB. *J Mol Biol* 1996,**261**:599-606.
270. Agostini I, Navarro JM, Bouhamdan M, Willetts K, Rey F, Spire B, *et al.* The HIV-1 Vpr co-activator induces a conformational change in TFIIB. *FEBS Lett* 1999,**450**:235-239.
271. Terada Y, Yasuda Y. Human immunodeficiency virus type 1 Vpr induces G2 checkpoint activation by interacting with the splicing factor SAP145. *Mol Cell Biol* 2006,**26**:8149-8158.
272. Hashizume C, Kuramitsu M, Zhang X, Kurosawa T, Kamata M, Aida Y. Human immunodeficiency virus type 1 Vpr interacts with spliceosomal protein SAP145 to mediate cellular pre-mRNA splicing inhibition. *Microbes Infect* 2007,**9**:490-497.
273. Goh WC, Manel N, Emerman M. The human immunodeficiency virus Vpr protein binds Cdc25C: implications for G2 arrest. *Virology* 2004,**318**:337-349.
274. Kino T, Gragerov A, Valentin A, Tsopanomihalou M, Ilyina-Gragerova G, Erwin-Cohen R, *et al.* Vpr protein of human immunodeficiency virus type 1 binds to 14-3-3 proteins and facilitates complex formation with Cdc25C: implications for cell cycle arrest. *J Virol* 2005,**79**:2780-2787.
275. Sherman MP, de Noronha CM, Heusch MI, Greene S, Greene WC. Nucleocytoplasmic shuttling by human immunodeficiency virus type 1 Vpr. *J Virol* 2001,**75**:1522-1532.
276. Schrofelbauer B, Hakata Y, Landau NR. HIV-1 Vpr function is mediated by interaction with the damage-specific DNA-binding protein DDB1. *Proc Natl Acad Sci U S A* 2007,**104**:4130-4135.
277. Sherman MP, de Noronha CM, Pearce D, Greene WC. Human immunodeficiency virus type 1 Vpr contains two leucine-rich helices that mediate glucocorticoid receptor coactivation independently of its effects on G(2) cell cycle arrest. *J Virol* 2000,**74**:8159-8165.
278. Muthumani K, Choo AY, Zong WX, Madesh M, Hwang DS, Premkumar A, *et al.* The HIV-1 Vpr and glucocorticoid receptor complex is a gain-of-function interaction that prevents the nuclear localization of PARP-1. *Nat Cell Biol* 2006,**8**:170-179.
279. Lai M, Zimmerman ES, Planelles V, Chen J. Activation of the ATR pathway by human immunodeficiency virus type 1 Vpr involves its direct binding to chromatin in vivo. *J Virol* 2005,**79**:15443-15451.
280. Stark LA, Hay RT. Human immunodeficiency virus type 1 (HIV-1) viral protein R (Vpr) interacts with Lys-tRNA synthetase: implications for priming of HIV-1 reverse transcription. *J Virol* 1998,**72**:3037-3044.
281. Wang L, Mukherjee S, Jia F, Narayan O, Zhao LJ. Interaction of virion protein Vpr of human immunodeficiency virus type 1 with cellular transcription factor Sp1 and trans-activation of viral long terminal repeat. *J Biol Chem* 1995,**270**:25564-25569.

282. Taneichi D, Iijima K, Doi A, Koyama T, Minemoto Y, Tokunaga K, *et al.* Identification of SNF2h, a chromatin-remodeling factor, as a novel binding protein of Vpr of human immunodeficiency virus type 1. *J Neuroimmune Pharmacol* 2011,**6**:177-187.
283. Godet AN, Guergnon J, Croset A, Cayla X, Falanga PB, Colle JH, *et al.* PP2A1 binding, cell transducing and apoptotic properties of Vpr(77-92): a new functional domain of HIV-1 Vpr proteins. *PLoS One* 2010,**5**:e13760.
284. Janoo A, Morrow PW, Tung HY. Activation of protein phosphatase-2A1 by HIV-1 Vpr cell death causing peptide in intact CD(4+) T cells and in vitro. *J Cell Biochem* 2005,**94**:816-825.
285. Sawaya BE, Khalili K, Gordon J, Taube R, Amini S. Cooperative interaction between HIV-1 regulatory proteins Tat and Vpr modulates transcription of the viral genome. *J Biol Chem* 2000,**275**:35209-35214.
286. Fouchier RA, Meyer BE, Simon JH, Fischer U, Albright AV, Gonzalez-Scarano F, *et al.* Interaction of the human immunodeficiency virus type 1 Vpr protein with the nuclear pore complex. *J Virol* 1998,**72**:6004-6013.
287. Cui J, Tungaturthi PK, Ayyavoo V, Ghafouri M, Ariga H, Khalili K, *et al.* The role of Vpr in the regulation of HIV-1 gene expression. *Cell Cycle* 2006,**5**:2626-2638.
288. Ramanathan MP, Curley E, 3rd, Su M, Chambers JA, Weiner DB. Carboxyl terminus of hVIP/mov34 is critical for HIV-1-Vpr interaction and glucocorticoid-mediated signaling. *J Biol Chem* 2002,**277**:47854-47860.
289. Mahalingam S, Ayyavoo V, Patel M, Kieber-Emmons T, Kao GD, Muschel RJ, *et al.* HIV-1 Vpr interacts with a human 34-kDa mov34 homologue, a cellular factor linked to the G2/M phase transition of the mammalian cell cycle. *Proc Natl Acad Sci U S A* 1998,**95**:3419-3424.
290. Hrecka K, Hao C, Gierszewska M, Swanson SK, Kesik-Brodacka M, Srivastava S, *et al.* Vpx relieves inhibition of HIV-1 infection of macrophages mediated by the SAMHD1 protein. *Nature* 2011,**474**:658-661.
291. Ahn J, Hao C, Yan J, DeLucia M, Mehrens J, Wang C, *et al.* HIV/simian immunodeficiency virus (SIV) accessory virulence factor Vpx loads the host cell restriction factor SAMHD1 onto the E3 ubiquitin ligase complex CRL4DCAF1. *J Biol Chem* 2012,**287**:12550-12558.
292. Laguette N, Rahm N, Sobhian B, Chable-Bessia C, Munch J, Snoeck J, *et al.* Evolutionary and functional analyses of the interaction between the myeloid restriction factor SAMHD1 and the lentiviral Vpx protein. *Cell Host Microbe* 2012,**11**:205-217.
293. Wei W, Guo H, Han X, Liu X, Zhou X, Zhang W, *et al.* A novel DCAF1-binding motif required for Vpx-mediated degradation of nuclear SAMHD1 and Vpr-induced G2 arrest. *Cell Microbiol* 2012,**14**:1745-1756.
294. Mueller SM, Jung R, Weiler S, Lang SM. Vpx proteins of SIVmac239 and HIV-2ROD interact with the cytoskeletal protein alpha-actinin 1. *J Gen Virol* 2004,**85**:3291-3303.
295. Cheng X, Belshan M, Ratner L. Hsp40 facilitates nuclear import of the human immunodeficiency virus type 2 Vpx-mediated preintegration complex. *J Virol* 2008,**82**:1229-1237.
296. Bergamaschi A, Ayinde D, David A, Le Rouzic E, Morel M, Collin G, *et al.* The human immunodeficiency virus type 2 Vpx protein usurps the CUL4A-DDB1DCAF1 ubiquitin ligase to overcome a postentry block in macrophage infection. *Journal of virology* 2009,**83**:4854-4860.
297. McCulley A, Ratner L. HIV-2 viral protein X (Vpx) ubiquitination is dispensable for ubiquitin ligase interaction and effects on macrophage infection. *Virology* 2012,**427**:67-75.
298. Berger A, Munk C, Schweizer M, Cichutek K, Schule S, Flory E. Interaction of Vpx and apolipoprotein B mRNA-editing catalytic polypeptide 3 family member A (APOBEC3A) correlates with efficient lentivirus infection of monocytes. *J Biol Chem* 2010,**285**:12248-12254.
299. Singhal PK, Rajendra Kumar P, Subba Rao MR, Mahalingam S. Nuclear export of simian immunodeficiency virus Vpx protein. *J Virol* 2006,**80**:12271-12282.
300. Rajendra Kumar P, Singhal PK, Subba Rao MR, Mahalingam S. Phosphorylation by MAPK regulates simian immunodeficiency virus Vpx protein nuclear import and virus infectivity. *J Biol Chem* 2005,**280**:8553-8563.
301. Pancio HA, Vander Heyden N, Kosuri K, Cresswell P, Ratner L. Interaction of human immunodeficiency virus type 2 Vpx and invariant chain. *J Virol* 2000,**74**:6168-6172.
302. Deng L, Ammosova T, Pumfery A, Kashanchi F, Nekhai S. HIV-1 Tat interaction with RNA polymerase II C-terminal domain (CTD) and a dynamic association with CDK2 induce CTD phosphorylation and transcription from HIV-1 promoter. *J Biol Chem* 2002,**277**:33922-33929.

303. Zhou C, Rana TM. A bimolecular mechanism of HIV-1 Tat protein interaction with RNA polymerase II transcription elongation complexes. *J Mol Biol* 2002,**320**:925-942.
304. Mavankal G, Ignatius Ou SH, Oliver H, Sigman D, Gaynor RB. Human immunodeficiency virus type 1 and 2 Tat proteins specifically interact with RNA polymerase II. *Proc Natl Acad Sci U S A* 1996,**93**:2089-2094.
305. Zhou M, Deng L, Kashanchi F, Brady JN, Shatkin AJ, Kumar A. The Tat/TAR-dependent phosphorylation of RNA polymerase II C-terminal domain stimulates cotranscriptional capping of HIV-1 mRNA. *Proc Natl Acad Sci U S A* 2003,**100**:12666-12671.
306. Herrmann CH, Rice AP. Lentivirus Tat proteins specifically associate with a cellular protein kinase, TAK, that hyperphosphorylates the carboxyl-terminal domain of the large subunit of RNA polymerase II: candidate for a Tat cofactor. *J Virol* 1995,**69**:1612-1620.
307. Wei P, Garber ME, Fang SM, Fischer WH, Jones KA. A novel CDK9-associated C-type cyclin interacts directly with HIV-1 Tat and mediates its high-affinity, loop-specific binding to TAR RNA. *Cell* 1998,**92**:451-462.
308. Bieniasz PD, Grdina TA, Bogerd HP, Cullen BR. Analysis of the effect of natural sequence variation in Tat and in cyclin T on the formation and RNA binding properties of Tat-cyclin T complexes. *J Virol* 1999,**73**:5777-5786.
309. Garber ME, Wei P, KewalRamani VN, Mayall TP, Herrmann CH, Rice AP, *et al.* The interaction between HIV-1 Tat and human cyclin T1 requires zinc and a critical cysteine residue that is not conserved in the murine CycT1 protein. *Genes Dev* 1998,**12**:3512-3527.
310. Garber ME, Mayall TP, Suess EM, Meisenhelder J, Thompson NE, Jones KA. CDK9 autophosphorylation regulates high-affinity binding of the human immunodeficiency virus type 1 tat-P-TEFb complex to TAR RNA. *Mol Cell Biol* 2000,**20**:6958-6969.
311. Tahirov TH, Babayeva ND, Varzavand K, Cooper JJ, Sedore SC, Price DH. Crystal structure of HIV-1 Tat complexed with human P-TEFb. *Nature* 2010,**465**:747-751.
312. Schulze-Gahmen U, Upton H, Birnberg A, Bao K, Chou S, Krogan NJ, *et al.* The AFF4 scaffold binds human P-TEFb adjacent to HIV Tat. *Elife* 2013,**2**:e00327.
313. Kiernan RE, Vanhulle C, Schiltz L, Adam E, Xiao H, Maudoux F, *et al.* HIV-1 tat transcriptional activity is regulated by acetylation. *EMBO J* 1999,**18**:6106-6118.
314. Ott M, Schnolzer M, Garnica J, Fischle W, Emiliani S, Rackwitz HR, *et al.* Acetylation of the HIV-1 Tat protein by p300 is important for its transcriptional activity. *Curr Biol* 1999,**9**:1489-1492.
315. Bres V, Tagami H, Peloponese JM, Loret E, Jeang KT, Nakatani Y, *et al.* Differential acetylation of Tat coordinates its interaction with the co-activators cyclin T1 and PCAF. *EMBO J* 2002,**21**:6811-6819.
316. Dorr A, Kiermer V, Pedal A, Rackwitz HR, Henklein P, Schubert U, *et al.* Transcriptional synergy between Tat and PCAF is dependent on the binding of acetylated Tat to the PCAF bromodomain. *EMBO J* 2002,**21**:2715-2723.
317. Hottiger MO, Nabel GJ. Interaction of human immunodeficiency virus type 1 Tat with the transcriptional coactivators p300 and CREB binding protein. *J Virol* 1998,**72**:8252-8256.
318. Mujtaba S, He Y, Zeng L, Farooq A, Carlson JE, Ott M, *et al.* Structural basis of lysine-acetylated HIV-1 Tat recognition by PCAF bromodomain. *Mol Cell* 2002,**9**:575-586.
319. Deng L, de la Fuente C, Fu P, Wang L, Donnelly R, Wade JD, *et al.* Acetylation of HIV-1 Tat by CBP/P300 increases transcription of integrated HIV-1 genome and enhances binding to core histones. *Virology* 2000,**277**:278-295.
320. Pantano S, Marcello A, Ferrari A, Gaudiosi D, Sabo A, Pellegrini V, *et al.* Insights on HIV-1 Tat:P/CAF bromodomain molecular recognition from in vivo experiments and molecular dynamics simulations. *Proteins* 2006,**62**:1062-1073.
321. Mahmoudi T, Parra M, Vries RG, Kauder SE, Verrijzer CP, Ott M, *et al.* The SWI/SNF chromatin-remodeling complex is a cofactor for Tat transactivation of the HIV promoter. *J Biol Chem* 2006,**281**:19960-19968.
322. Benkirane M, Chun RF, Xiao H, Ogryzko VV, Howard BH, Nakatani Y, *et al.* Activation of integrated provirus requires histone acetyltransferase. p300 and P/CAF are coactivators for HIV-1 Tat. *J Biol Chem* 1998,**273**:24898-24905.
323. Marzio G, Tyagi M, Gutierrez MI, Giacca M. HIV-1 tat transactivator recruits p300 and CREB-binding protein histone acetyltransferases to the viral promoter. *Proc Natl Acad Sci U S A* 1998,**95**:13519-13524.
324. Gabizon R, Mor M, Rosenberg MM, Britan L, Hayouka Z, Kotler M, *et al.* Using peptides to study the interaction between the p53 tetramerization domain and HIV-1 Tat. *Biopolymers* 2008,**90**:105-116.

325. Longo F, Marchetti MA, Castagnoli L, Battaglia PA, Gigliani F. A novel approach to protein-protein interaction: complex formation between the p53 tumor suppressor and the HIV Tat proteins. *Biochem Biophys Res Commun* 1995,**206**:326-334.
326. Huang X, Seifert U, Salzmann U, Henklein P, Preissner R, Henke W, *et al.* The RTP site shared by the HIV-1 Tat protein and the 11S regulator subunit alpha is crucial for their effects on proteasome function including antigen processing. *J Mol Biol* 2002,**323**:771-782.
327. Fiume G, Vecchio E, De Laurentiis A, Trimboli F, Palmieri C, Pisano A, *et al.* Human immunodeficiency virus-1 Tat activates NF-kappaB via physical interaction with IkappaB-alpha and p65. *Nucleic Acids Res* 2012,**40**:3548-3562.
328. Puca A, Fiume G, Palmieri C, Trimboli F, Olimpico F, Scala G, *et al.* IkappaB-alpha represses the transcriptional activity of the HIV-1 Tat transactivator by promoting its nuclear export. *J Biol Chem* 2007,**282**:37146-37157.
329. Berro R, Kehn K, de la Fuente C, Pumfery A, Adair R, Wade J, *et al.* Acetylated Tat regulates human immunodeficiency virus type 1 splicing through its interaction with the splicing regulator p32. *J Virol* 2006,**80**:3189-3204.
330. Yu L, Loewenstein PM, Zhang Z, Green M. In vitro interaction of the human immunodeficiency virus type 1 Tat transactivator and the general transcription factor TFIIB with the cellular protein TAP. *J Virol* 1995,**69**:3017-3023.
331. Yu L, Zhang Z, Loewenstein PM, Desai K, Tang Q, Mao D, *et al.* Molecular cloning and characterization of a cellular protein that interacts with the human immunodeficiency virus type 1 Tat transactivator and encodes a strong transcriptional activation domain. *J Virol* 1995,**69**:3007-3016.
332. Luo Y, Yu H, Peterlin BM. Cellular protein modulates effects of human immunodeficiency virus type 1 Rev. *J Virol* 1994,**68**:3850-3856.
333. Rom S, Pacifici M, Passiatore G, Aprea S, Waligorska A, Del Valle L, *et al.* HIV-1 Tat binds to SH3 domains: cellular and viral outcome of Tat/Grb2 interaction. *Biochim Biophys Acta* 2011,**1813**:1836-1844.
334. Ambrosino C, Palmieri C, Puca A, Trimboli F, Schiavone M, Olimpico F, *et al.* Physical and functional interaction of HIV-1 Tat with E2F-4, a transcriptional regulator of mammalian cell cycle. *J Biol Chem* 2002,**277**:31448-31458.
335. Kwon HS, Brent MM, Getachew R, Jayakumar P, Chen LF, Schnolzer M, *et al.* Human immunodeficiency virus type 1 Tat protein inhibits the SIRT1 deacetylase and induces T cell hyperactivation. *Cell Host Microbe* 2008,**3**:158-167.
336. Pagans S, Pedal A, North BJ, Kaehlcke K, Marshall BL, Dorr A, *et al.* SIRT1 regulates HIV transcription via Tat deacetylation. *PLoS Biol* 2005,**3**:e41.
337. Fridell RA, Harding LS, Bogerd HP, Cullen BR. Identification of a novel human zinc finger protein that specifically interacts with the activation domain of lentiviral Tat proteins. *Virology* 1995,**209**:347-357.
338. Ariumi Y, Serhan F, Turelli P, Telenti A, Trono D. The integrase interactor 1 (INI1) proteins facilitate Tat-mediated human immunodeficiency virus type 1 transcription. *Retrovirology* 2006,**3**:47.
339. He L, Liu H, Tang L. SWI/SNF chromatin remodeling complex: a new cofactor in reprogramming. *Stem Cell Rev* 2012,**8**:128-136.
340. Agbottah E, Deng L, Dannenberg LO, Pumfery A, Kashanchi F. Effect of SWI/SNF chromatin remodeling complex on HIV-1 Tat activated transcription. *Retrovirology* 2006,**3**:48.
341. Epie N, Ammosova T, Sapir T, Voloshin Y, Lane WS, Turner W, *et al.* HIV-1 Tat interacts with LIS1 protein. *Retrovirology* 2005,**2**:6.
342. Weissman JD, Brown JA, Howcroft TK, Hwang J, Chawla A, Roche PA, *et al.* HIV-1 tat binds TAFII250 and represses TAFII250-dependent transcription of major histocompatibility class I genes. *Proc Natl Acad Sci U S A* 1998,**95**:11601-11606.
343. Weissman JD, Hwang JR, Singer DS. Extensive interactions between HIV TAT and TAF(II)250. *Biochim Biophys Acta* 2001,**1546**:156-163.
344. Chiu YL, Coronel E, Ho CK, Shuman S, Rana TM. HIV-1 Tat protein interacts with mammalian capping enzyme and stimulates capping of TAR RNA. *J Biol Chem* 2001,**276**:12959-12966.
345. Col E, Caron C, Seigneurin-Berny D, Gracia J, Favier A, Khochbin S. The histone acetyltransferase, hGCN5, interacts with and acetylates the HIV transactivator, Tat. *J Biol Chem* 2001,**276**:28179-28184.

346. Wortman MJ, Krachmarov CP, Kim JH, Gordon RG, Chepenik LG, Brady JN, *et al.* Interaction of HIV-1 Tat with Puralpha in nuclei of human glial cells: characterization of RNA-mediated protein-protein binding. *J Cell Biochem* 2000,**77**:65-74.
347. Peruzzi F, Gordon J, Darbinian N, Amini S. Tat-induced deregulation of neuronal differentiation and survival by nerve growth factor pathway. *J Neurovirol* 2002,**8 Suppl 2**:91-96.
348. Gallia GL, Darbinian N, Tretiakova A, Ansari SA, Rappaport J, Brady J, *et al.* Association of HIV-1 Tat with the cellular protein, Puralpha, is mediated by RNA. *Proc Natl Acad Sci U S A* 1999,**96**:11572-11577.
349. Ansari SA, Safak M, Gallia GL, Sawaya BE, Amini S, Khalili K. Interaction of YB-1 with human immunodeficiency virus type 1 Tat and TAR RNA modulates viral promoter activity. *J Gen Virol* 1999,**80 (Pt 10)**:2629-2638.
350. Bennasser Y, Jeang KT. HIV-1 Tat interaction with Dicer: requirement for RNA. *Retrovirology* 2006,**3**:95.
351. Rohr O, Lecestre D, Chasserot-Golaz S, Marban C, Avram D, Aunis D, *et al.* Recruitment of Tat to heterochromatin protein HP1 via interaction with CTIP2 inhibits human immunodeficiency virus type 1 replication in microglial cells. *J Virol* 2003,**77**:5415-5427.
352. Macian F, Rao A. Reciprocal modulatory interaction between human immunodeficiency virus type 1 Tat and transcription factor NFAT1. *Mol Cell Biol* 1999,**19**:3645-3653.
353. Hidalgo-Estevez AM, Gonzalez E, Punzon C, Fresno M. Human immunodeficiency virus type 1 Tat increases cooperation between AP-1 and NFAT transcription factors in T cells. *J Gen Virol* 2006,**87**:1603-1612.
354. Vardabasso C, Manganaro L, Lusic M, Marcello A, Giacca M. The histone chaperone protein Nucleosome Assembly Protein-1 (hNAP-1) binds HIV-1 Tat and promotes viral transcription. *Retrovirology* 2008,**5**:8.
355. Xiao H, Neuveut C, Benkirane M, Jeang KT. Interaction of the second coding exon of Tat with human EF-1 delta delineates a mechanism for HIV-1-mediated shut-off of host mRNA translation. *Biochem Biophys Res Commun* 1998,**244**:384-389.
356. Ammosova T, Jerebtsova M, Beullens M, Lesage B, Jackson A, Kashanchi F, *et al.* Nuclear targeting of protein phosphatase-1 by HIV-1 Tat protein. *J Biol Chem* 2005,**280**:36364-36371.
357. Kino T, Slobodskaya O, Pavlakis GN, Chrousos GP. Nuclear receptor coactivator p160 proteins enhance the HIV-1 long terminal repeat promoter by bridging promoter-bound factors and the Tat-P-TEFb complex. *J Biol Chem* 2002,**277**:2396-2405.
358. Li W, Huang Y, Reid R, Steiner J, Malpica-Llanos T, Darden TA, *et al.* NMDA receptor activation by HIV-Tat protein is clade dependent. *J Neurosci* 2008,**28**:12190-12198.
359. Prendergast MA, Rogers DT, Mulholland PJ, Littleton JM, Wilkins LH, Jr., Self RL, *et al.* Neurotoxic effects of the human immunodeficiency virus type-1 transcription factor Tat require function of a polyamine sensitive-site on the N-methyl-D-aspartate receptor. *Brain Res* 2002,**954**:300-307.
360. Kashanchi F, Piras G, Radonovich MF, Duvall JF, Fattaey A, Chiang CM, *et al.* Direct interaction of human TFIID with the HIV-1 transactivator tat. *Nature* 1994,**367**:295-299.
361. Veschambre P, Simard P, Jalinot P. Evidence for functional interaction between the HIV-1 Tat transactivator and the TATA box binding protein in vivo. *J Mol Biol* 1995,**250**:169-180.
362. Veschambre P, Roisin A, Jalinot P. Biochemical and functional interaction of the human immunodeficiency virus type 1 Tat transactivator with the general transcription factor TFIIB. *J Gen Virol* 1997,**78 (Pt 9)**:2235-2245.
363. Parada CA, Roeder RG. Enhanced processivity of RNA polymerase II triggered by Tat-induced phosphorylation of its carboxy-terminal domain. *Nature* 1996,**384**:375-378.
364. Zhou M, Nekhai S, Bharucha DC, Kumar A, Ge H, Price DH, *et al.* TFIIH inhibits CDK9 phosphorylation during human immunodeficiency virus type 1 transcription. *J Biol Chem* 2001,**276**:44633-44640.
365. McQueen P, Donald LJ, Vo TN, Nguyen DH, Griffiths H, Shojania S, *et al.* Tat peptide-calmodulin binding studies and bioinformatics of HIV-1 protein-calmodulin interactions. *Proteins* 2011,**79**:2233-2246.
366. Haij NB, Leghmari K, Planes R, Thieblemont N, Bahraoui E. HIV-1 Tat protein binds to TLR4-MD2 and signals to induce TNF-alpha and IL-10. *Retrovirology* 2013,**10**:123.
367. Treand C, du Chene I, Bres V, Kiernan R, Benarous R, Benkirane M, *et al.* Requirement for SWI/SNF chromatin-remodeling complex in Tat-mediated activation of the HIV-1 promoter. *EMBO J* 2006,**25**:1690-1699.

368. Urbinati C, Bugatti A, Giacca M, Schlaepfer D, Presta M, Rusnati M. alpha(v)beta3-integrin-dependent activation of focal adhesion kinase mediates NF-kappaB activation and motogenic activity by HIV-1 Tat in endothelial cells. *J Cell Sci* 2005,**118**:3949-3958.
369. Barillari G, Sgadari C, Fiorelli V, Samaniego F, Colombini S, Manzari V, *et al.* The Tat protein of human immunodeficiency virus type-1 promotes vascular cell growth and locomotion by engaging the alpha5beta1 and alphavbeta3 integrins and by mobilizing sequestered basic fibroblast growth factor. *Blood* 1999,**94**:663-672.
370. Vogel BE, Lee SJ, Hildebrand A, Craig W, Pierschbacher MD, Wong-Staal F, *et al.* A novel integrin specificity exemplified by binding of the alpha v beta 5 integrin to the basic domain of the HIV Tat protein and vitronectin. *J Cell Biol* 1993,**121**:461-468.
371. Kamine J, Elangovan B, Subramanian T, Coleman D, Chinnadurai G. Identification of a cellular protein that specifically interacts with the essential cysteine region of the HIV-1 Tat transactivator. *Virology* 1996,**216**:357-366.
372. Xiao H, Tao Y, Greenblatt J, Roeder RG. A cofactor, TIP30, specifically enhances HIV-1 Tat-activated transcription. *Proc Natl Acad Sci U S A* 1998,**95**:2146-2151.
373. El Omari K, Bird LE, Nichols CE, Ren J, Stammers DK. Crystal structure of CC3 (TIP30): implications for its role as a tumor suppressor. *J Biol Chem* 2005,**280**:18229-18236.
374. Abraham S, Sweet T, Sawaya BE, Rappaport J, Khalili K, Amini S. Cooperative interaction of C/EBP beta and Tat modulates MCP-1 gene transcription in astrocytes. *J Neuroimmunol* 2005,**160**:219-227.
375. Albini A, Soldi R, Giunciuglio D, Giraudo E, Benelli R, Primo L, *et al.* The angiogenesis induced by HIV-1 tat protein is mediated by the Flk-1/KDR receptor on vascular endothelial cells. *Nat Med* 1996,**2**:1371-1375.
376. Albini A, Ferrini S, Benelli R, Sforzini S, Giunciuglio D, Aluigi MG, *et al.* HIV-1 Tat protein mimicry of chemokines. *Proc Natl Acad Sci U S A* 1998,**95**:13153-13158.
377. Ghezzi S, Noonan DM, Aluigi MG, Vallanti G, Cota M, Benelli R, *et al.* Inhibition of CXCR4-dependent HIV-1 infection by extracellular HIV-1 Tat. *Biochem Biophys Res Commun* 2000,**270**:992-996.
378. Liu Y, Jones M, Hingtgen CM, Bu G, Laribee N, Tanzi RE, *et al.* Uptake of HIV-1 tat protein mediated by low-density lipoprotein receptor-related protein disrupts the neuronal metabolic balance of the receptor ligands. *Nat Med* 2000,**6**:1380-1387.
379. Chen D, Wang M, Zhou S, Zhou Q. HIV-1 Tat targets microtubules to induce apoptosis, a process promoted by the pro-apoptotic Bcl-2 relative Bim. *EMBO J* 2002,**21**:6801-6810.
380. Ziegler A, Seelig J. Interaction of the protein transduction domain of HIV-1 TAT with heparan sulfate: binding mechanism and thermodynamic parameters. *Biophys J* 2004,**86**:254-263.
381. De Francesco MA, Baronio M, Poiesi C. HIV-1 p17 matrix protein interacts with heparan sulfate side chain of CD44v3, syndecan-2, and syndecan-4 proteoglycans expressed on human activated CD4+ T cells affecting tumor necrosis factor alpha and interleukin 2 production. *J Biol Chem* 2011,**286**:19541-19548.
382. Mitola S, Soldi R, Zanon I, Barra L, Gutierrez MI, Berkhout B, *et al.* Identification of specific molecular structures of human immunodeficiency virus type 1 Tat relevant for its biological effects on vascular endothelial cells. *J Virol* 2000,**74**:344-353.
383. Watson K, Gooderham NJ, Davies DS, Edwards RJ. Interaction of the transactivating protein HIV-1 tat with sulphated polysaccharides. *Biochem Pharmacol* 1999,**57**:775-783.
384. Zhu J, Mactutus CF, Wallace DR, Booze RM. HIV-1 Tat protein-induced rapid and reversible decrease in [3H]dopamine uptake: dissociation of [3H]dopamine uptake and [3H]2beta-carbomethoxy-3-beta-(4-fluorophenyl)tropane (WIN 35,428) binding in rat striatal synaptosomes. *J Pharmacol Exp Ther* 2009,**329**:1071-1083.
385. Ekokoski E, Aitio O, Tornquist K, Yli-Kauhaluoma J, Tuominen RK. HIV-1 Tat-peptide inhibits protein kinase C and protein kinase A through substrate competition. *Eur J Pharm Sci* 2010,**40**:404-411.
386. Holmes AM. In vitro phosphorylation of human immunodeficiency virus type 1 Tat protein by protein kinase C: evidence for the phosphorylation of amino acid residue serine-46. *Arch Biochem Biophys* 1996,**335**:8-12.
387. Cai R, Carpick B, Chun RF, Jeang KT, Williams BR. HIV-I TAT inhibits PKR activity by both RNA-dependent and RNA-independent mechanisms. *Arch Biochem Biophys* 2000,**373**:361-367.

388. Brand SR, Kobayashi R, Mathews MB. The Tat protein of human immunodeficiency virus type 1 is a substrate and inhibitor of the interferon-induced, virally activated protein kinase, PKR. *J Biol Chem* 1997,**272**:8388-8395.
389. Bres V, Kiernan RE, Linares LK, Chable-Bessia C, Plechakova O, Treand C, *et al.* A non-proteolytic role for ubiquitin in Tat-mediated transactivation of the HIV-1 promoter. *Nat Cell Biol* 2003,**5**:754-761.
390. Tikhonov I, Ruckwardt TJ, Berg S, Hatfield GS, David Pauza C. Furin cleavage of the HIV-1 Tat protein. *FEBS Lett* 2004,**565**:89-92.
391. Mediouni S, Watkins JD, Pierres M, Bole A, Loret EP, Baillat G. A monoclonal antibody directed against a conformational epitope of the HIV-1 trans-activator (Tat) protein neutralizes cross-clade. *J Biol Chem* 2012,**287**:11942-11950.
392. Serriere J, Dugua JM, Bossus M, Verrier B, Haser R, Gouet P, *et al.* Fab'-induced folding of antigenic N-terminal peptides from intrinsically disordered HIV-1 Tat revealed by X-ray crystallography. *J Mol Biol* 2011,**405**:33-42.
393. Sharma A, Awasthi S, Harrod CK, Matlock EF, Khan S, Xu L, *et al.* The Werner syndrome helicase is a cofactor for HIV-1 long terminal repeat transactivation and retroviral replication. *J Biol Chem* 2007,**282**:12048-12057.
394. Lai MC, Wang SW, Cheng L, Tarn WY, Tsai SJ, Sun HS. Human DDX3 interacts with the HIV-1 Tat protein to facilitate viral mRNA translation. *PLoS One* 2013,**8**:e68665.
395. Yasuda-Inoue M, Kuroki M, Ariumi Y. DDX3 RNA helicase is required for HIV-1 Tat function. *Biochem Biophys Res Commun* 2013,**441**:607-611.
396. Papandreou MJ, Barbouche R, Guieu R, Rivera S, Fantini J, Khrestchatisky M, *et al.* Mapping of domains on HIV envelope protein mediating association with calnexin and protein-disulfide isomerase. *J Biol Chem* 2010,**285**:13788-13796.
397. Wu Z, Golub E, Abrams WR, Malamud D. gp340 (SAG) binds to the V3 sequence of gp120 important for chemokine receptor interaction. *AIDS Res Hum Retroviruses* 2004,**20**:600-607.
398. Stoddard E, Cannon G, Ni H, Kariko K, Capodici J, Malamud D, *et al.* gp340 expressed on human genital epithelia binds HIV-1 envelope protein and facilitates viral transmission. *J Immunol* 2007,**179**:3126-3132.
399. Chu Y, Li J, Wu X, Hua Z, Wu Z. Identification of human immunodeficiency virus type 1 (HIV-1) gp120-binding sites on scavenger receptor cysteine rich 1 (SRCR1) domain of gp340. *J Biomed Sci* 2013,**20**:44.
400. Bowman MR, MacFerrin KD, Schreiber SL, Burakoff SJ. Identification and structural analysis of residues in the V1 region of CD4 involved in interaction with human immunodeficiency virus envelope glycoprotein gp120 and class II major histocompatibility complex molecules. *Proc Natl Acad Sci U S A* 1990,**87**:9052-9056.
401. Bour S, Geleziunas R, Wainberg MA. The human immunodeficiency virus type 1 (HIV-1) CD4 receptor and its central role in promotion of HIV-1 infection. *Microbiol Rev* 1995,**59**:63-93.
402. Tran EE, Borgnia MJ, Kuybeda O, Schauder DM, Bartesaghi A, Frank GA, *et al.* Structural mechanism of trimeric HIV-1 envelope glycoprotein activation. *PLoS Pathog* 2012,**8**:e1002797.
403. Zhou T, Georgiev I, Wu X, Yang ZY, Dai K, Finzi A, *et al.* Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science* 2010,**329**:811-817.
404. Huang CC, Lam SN, Acharya P, Tang M, Xiang SH, Hussan SS, *et al.* Structures of the CCR5 N terminus and of a tyrosine-sulfated antibody with HIV-1 gp120 and CD4. *Science* 2007,**317**:1930-1934.
405. Grupping K, Selhorst P, Michiels J, Vereecken K, Heyndrickx L, Kessler P, *et al.* MiniCD4 protein resistance mutations affect binding to the HIV-1 gp120 CD4 binding site and decrease entry efficiency. *Retrovirology* 2012,**9**:36.
406. Lu K, Heng X, Summers MF. Structural determinants and mechanism of HIV-1 genome packaging. *J Mol Biol* 2011,**410**:609-633.
407. Wu X, Zhou T, O'Dell S, Wyatt RT, Kwong PD, Mascola JR. Mechanism of human immunodeficiency virus type 1 resistance to monoclonal antibody B12 that effectively targets the site of CD4 attachment. *J Virol* 2009,**83**:10892-10907.
408. Lin G, Baribaud F, Romano J, Doms RW, Hoxie JA. Identification of gp120 binding sites on CXCR4 by using CD4-independent human immunodeficiency virus type 2 Env proteins. *J Virol* 2003,**77**:931-942.

409. Basmaciogullari S, Babcock GJ, Van Ryk D, Wojtowicz W, Sodroski J. Identification of conserved and variable structures in the human immunodeficiency virus gp120 glycoprotein of importance for CXCR4 binding. *J Virol* 2002,**76**:10791-10800.
410. Pollakis G, Kang S, Kliphuis A, Chalaby MI, Goudsmit J, Paxton WA. N-linked glycosylation of the HIV type-1 gp120 envelope glycoprotein as a major determinant of CCR5 and CXCR4 coreceptor utilization. *J Biol Chem* 2001,**276**:13433-13441.
411. Huang CC, Tang M, Zhang MY, Majeed S, Montabana E, Stanfield RL, *et al.* Structure of a V3-containing HIV-1 gp120 core. *Science* 2005,**310**:1025-1028.
412. Choe H, Farzan M, Sun Y, Sullivan N, Rollins B, Ponath PD, *et al.* The beta-chemokine receptors CCR3 and CCR5 facilitate infection by primary HIV-1 isolates. *Cell* 1996,**85**:1135-1148.
413. Mahajan SD, Aalinkeel R, Reynolds JL, Nair BB, Fernandez SF, Schwartz SA, *et al.* Morphine exacerbates HIV-1 viral protein gp120 induced modulation of chemokine gene expression in U373 astrocytoma cells. *Curr HIV Res* 2005,**3**:277-288.
414. Endrich MM, Gehring H. The V3 loop of human immunodeficiency virus type-1 envelope protein is a high-affinity ligand for immunophilins present in human blood. *Eur J Biochem* 1998,**252**:441-446.
415. Bosch V, Pawlita M. Mutational analysis of the human immunodeficiency virus type 1 env gene product proteolytic cleavage site. *J Virol* 1990,**64**:2337-2344.
416. Morozov VA, Morozov AV, Lagaye S. Short communication: Simultaneous substitutions of V38M and N43T-N44K in the gp41 heptad repeat 1 (HR1) disrupt HIV type 1 gPr160 endoproteolytic cleavage (*). *AIDS Res Hum Retroviruses* 2010,**26**:73-77.
417. Kibler KV, Miyazato A, Yedavalli VS, Dayton AI, Jacobs BL, Dapolito G, *et al.* Polyarginine inhibits gp160 processing by furin and suppresses productive human immunodeficiency virus type 1 infection. *J Biol Chem* 2004,**279**:49055-49063.
418. Crublet E, Andrieu JP, Vives RR, Lortat-Jacob H. The HIV-1 envelope glycoprotein gp120 features four heparan sulfate binding domains, including the co-receptor binding site. *J Biol Chem* 2008,**283**:15193-15200.
419. Witvrouw M, Fikkert V, Hantson A, Pannecouque C, O'Keefe B R, McMahon J, *et al.* Resistance of human immunodeficiency virus type 1 to the high-mannose binding agents cyanovirin N and concanavalin A. *J Virol* 2005,**79**:7777-7784.
420. Alexandre KB, Gray ES, Pantophlet R, Moore PL, McMahon JB, Chakauya E, *et al.* Binding of the mannose-specific lectin, griffithsin, to HIV-1 gp120 exposes the CD4-binding site. *J Virol* 2011,**85**:9039-9050.
421. Alexandre KB, Moore PL, Nonyane M, Gray ES, Ranchobe N, Chakauya E, *et al.* Mechanisms of HIV-1 subtype C resistance to GRFT, CV-N and SVN. *Virology* 2013,**446**:66-76.
422. Huang X, Jin W, Griffin GE, Shattock RJ, Hu Q. Removal of two high-mannose N-linked glycans on gp120 renders human immunodeficiency virus 1 largely resistant to the carbohydrate-binding agent griffithsin. *J Gen Virol* 2011,**92**:2367-2373.
423. Xue J, Gao Y, Hoorelbeke B, Kagiampakis I, Zhao B, Demeler B, *et al.* The role of individual carbohydrate-binding sites in the function of the potent anti-HIV lectin griffithsin. *Mol Pharm* 2012,**9**:2613-2625.
424. Alexandre KB, Gray ES, Lambson BE, Moore PL, Choge IA, Mlisana K, *et al.* Mannose-rich glycosylation patterns on HIV-1 subtype C gp120 and sensitivity to the lectins, Griffithsin, Cyanovirin-N and Scytovirin. *Virology* 2010,**402**:187-196.
425. Moulaei T, Shenoy SR, Giomarelli B, Thomas C, McMahon JB, Dauter Z, *et al.* Monomerization of viral entry inhibitor griffithsin elucidates the relationship between multivalent binding to carbohydrates and anti-HIV activity. *Structure* 2010,**18**:1104-1115.
426. Hong PW, Flummerfelt KB, de Parseval A, Gurney K, Elder JH, Lee B. Human immunodeficiency virus envelope (gp120) binding to DC-SIGN and primary dendritic cells is carbohydrate dependent but does not involve 2G12 or cyanovirin binding sites: implications for structural analyses of gp120-DC-SIGN binding. *J Virol* 2002,**76**:12855-12865.
427. Bokesch HR, O'Keefe BR, McKee TC, Pannell LK, Patterson GM, Gardella RS, *et al.* A potent novel anti-HIV protein from the cultured cyanobacterium Scytonema varium. *Biochemistry* 2003,**42**:2578-2584.
428. Chang LC, Bewley CA. Potent inhibition of HIV-1 fusion by cyanovirin-N requires only a single high affinity carbohydrate binding site: characterization of low affinity carbohydrate binding site knockout mutants. *J Mol Biol* 2002,**318**:1-8.

429. O'Keefe BR, Shenoy SR, Xie D, Zhang W, Muschik JM, Currens MJ, *et al.* Analysis of the interaction between the HIV-inactivating protein cyanovirin-N and soluble forms of the envelope glycoproteins gp120 and gp41. *Mol Pharmacol* 2000,**58**:982-992.
430. Barrientos LG, Louis JM, Ratner DM, Seeberger PH, Gronenborn AM. Solution structure of a circular-permuted variant of the potent HIV-inactivating protein cyanovirin-N: structural basis for protein stability and oligosaccharide interaction. *J Mol Biol* 2003,**325**:211-223.
431. Fromme R, Katiliene Z, Giomarelli B, Bogani F, Mc Mahon J, Mori T, *et al.* A monovalent mutant of cyanovirin-N provides insight into the role of multiple interactions with gp120 for antiviral activity. *Biochemistry* 2007,**46**:9199-9207.
432. Binley JM, Ngo-Abdalla S, Moore P, Bobardt M, Chatterji U, Gallay P, *et al.* Inhibition of HIV Env binding to cellular receptors by monoclonal antibody 2G12 as probed by Fc-tagged gp120. *Retrovirology* 2006,**3**:39.
433. Scanlan CN, Pantophlet R, Wormald MR, Ollmann Saphire E, Stanfield R, Wilson IA, *et al.* The broadly neutralizing anti-human immunodeficiency virus type 1 antibody 2G12 recognizes a cluster of alpha1-->2 mannose residues on the outer face of gp120. *J Virol* 2002,**76**:7306-7321.
434. Calarese DA, Lee HK, Huang CY, Best MD, Astronomo RD, Stanfield RL, *et al.* Dissection of the carbohydrate specificity of the broadly neutralizing anti-HIV-1 antibody 2G12. *Proc Natl Acad Sci U S A* 2005,**102**:13372-13377.
435. Trkola A, Purtscher M, Muster T, Ballaun C, Buchacher A, Sullivan N, *et al.* Human monoclonal antibody 2G12 defines a distinctive neutralization epitope on the gp120 glycoprotein of human immunodeficiency virus type 1. *J Virol* 1996,**70**:1100-1108.
436. Doores KJ, Fulton Z, Huber M, Wilson IA, Burton DR. Antibody 2G12 recognizes di-mannose equivalently in domain- and nondomain-exchanged forms but only binds the HIV-1 glycan shield if domain exchanged. *J Virol* 2010,**84**:10690-10699.
437. Calarese DA, Scanlan CN, Zwick MB, Deechongkit S, Mimura Y, Kunert R, *et al.* Antibody domain exchange is an immunological solution to carbohydrate cluster recognition. *Science* 2003,**300**:2065-2071.
438. Saphire EO, Parren PW, Pantophlet R, Zwick MB, Morris GM, Rudd PM, *et al.* Crystal structure of a neutralizing human IGG against HIV-1: a template for vaccine design. *Science* 2001,**293**:1155-1159.
439. Zhou T, Xu L, Dey B, Hessell AJ, Van Ryk D, Xiang SH, *et al.* Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Nature* 2007,**445**:732-737.
440. Kwong PD, Doyle ML, Casper DJ, Cicala C, Leavitt SA, Majeed S, *et al.* HIV-1 evades antibody-mediated neutralization through conformational masking of receptor-binding sites. *Nature* 2002,**420**:678-682.
441. Gorny MK, VanCott TC, Williams C, Revesz K, Zolla-Pazner S. Effects of oligomerization on the epitopes of the human immunodeficiency virus type 1 envelope glycoproteins. *Virology* 2000,**267**:220-228.
442. Kwong PD, Wyatt R, Robinson J, Sweet RW, Sodroski J, Hendrickson WA. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* 1998,**393**:648-659.
443. Huang CC, Venturi M, Majeed S, Moore MJ, Phogat S, Zhang MY, *et al.* Structural basis of tyrosine sulfation and VH-gene usage in antibodies that recognize the HIV type 1 coreceptor-binding site on gp120. *Proc Natl Acad Sci U S A* 2004,**101**:2706-2711.
444. Thali M, Olshevsky U, Furman C, Gabuzda D, Posner M, Sodroski J. Characterization of a discontinuous human immunodeficiency virus type 1 gp120 epitope recognized by a broadly reactive neutralizing human monoclonal antibody. *J Virol* 1991,**65**:6188-6193.
445. Chen L, Kwon YD, Zhou T, Wu X, O'Dell S, Cavacini L, *et al.* Structural basis of immune evasion at the site of CD4 attachment on HIV-1 gp120. *Science* 2009,**326**:1123-1127.
446. Xiang SH, Wang L, Abreu M, Huang CC, Kwong PD, Rosenberg E, *et al.* Epitope mapping and characterization of a novel CD4-induced human monoclonal antibody capable of neutralizing primary HIV-1 strains. *Virology* 2003,**315**:124-134.
447. Dorfman T, Moore MJ, Guth AC, Choe H, Farzan M. A tyrosine-sulfated peptide derived from the heavy-chain CDR3 region of an HIV-1-neutralizing antibody binds gp120 and inhibits HIV-1 infection. *J Biol Chem* 2006,**281**:28529-28535.
448. Gorny MK, Moore JP, Conley AJ, Karwowska S, Sodroski J, Williams C, *et al.* Human anti-V2 monoclonal antibody that neutralizes primary but not laboratory isolates of human immunodeficiency virus type 1. *J Virol* 1994,**68**:8312-8320.

449. West AP, Jr., Scharf L, Horwitz J, Klein F, Nussenzweig MC, Bjorkman PJ. Computational analysis of anti-HIV-1 antibody neutralization panel data to identify potential functional epitope residues. *Proc Natl Acad Sci U S A* 2013,**110**:10598-10603.
450. Murphy MK, Yue L, Pan R, Boliar S, Sethi A, Tian J, *et al.* Viral escape from neutralizing antibodies in early subtype A HIV-1 infection drives an increase in autologous neutralization breadth. *PLoS Pathog* 2013,**9**:e1003173.
451. Zhang MY, Shu Y, Rudolph D, Prabakaran P, Labrijn AF, Zwick MB, *et al.* Improved breadth and potency of an HIV-1-neutralizing human single-chain antibody by random mutagenesis and sequential antigen panning. *J Mol Biol* 2004,**335**:209-219.
452. Zhang MY, Yuan T, Li J, Rosa Borges A, Watkins JD, Guenaga J, *et al.* Identification and characterization of a broadly cross-reactive HIV-1 human monoclonal antibody that binds to both gp120 and gp41. *PLoS One* 2012,**7**:e44241.
453. Sanders RW, Venturi M, Schiffner L, Kalyanaraman R, Katinger H, Lloyd KO, *et al.* The mannose-dependent epitope for neutralizing antibody 2G12 on human immunodeficiency virus type 1 glycoprotein gp120. *J Virol* 2002,**76**:7293-7305.
454. Moore PL, Gray ES, Sheward D, Madiga M, Ranchobe N, Lai Z, *et al.* Potent and broad neutralization of HIV-1 subtype C by plasma antibodies targeting a quaternary epitope including residues in the V2 loop. *J Virol* 2011,**85**:3128-3141.
455. Pantophlet R, Aguilar-Sino RO, Wrin T, Cavacini LA, Burton DR. Analysis of the neutralization breadth of the anti-V3 antibody F425-B4e8 and re-assessment of its epitope fine specificity by scanning mutagenesis. *Virology* 2007,**364**:441-453.
456. Bell CH, Pantophlet R, Schiefner A, Cavacini LA, Stanfield RL, Burton DR, *et al.* Structure of antibody F425-B4e8 in complex with a V3 peptide reveals a new binding mode for HIV-1 neutralization. *J Mol Biol* 2008,**375**:969-978.
457. Klein F, Gaebler C, Mouquet H, Sather DN, Lehmann C, Scheid JF, *et al.* Broad neutralization by a combination of antibodies recognizing the CD4 binding site and a new conformational epitope on the HIV-1 envelope protein. *J Exp Med* 2012,**209**:1469-1479.
458. Pejchal R, Walker LM, Stanfield RL, Phogat SK, Koff WC, Poignard P, *et al.* Structure and function of broadly reactive antibody PG16 reveal an H3 subdomain that mediates potent neutralization of HIV-1. *Proc Natl Acad Sci U S A* 2010,**107**:11483-11488.
459. McLellan JS, Pancera M, Carrico C, Gorman J, Julien JP, Khayat R, *et al.* Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9. *Nature* 2011,**480**:336-343.
460. Doria-Rose NA, Georgiev I, O'Dell S, Chuang GY, Staupe RP, McLellan JS, *et al.* A short segment of the HIV-1 gp120 V1/V2 region is a major determinant of resistance to V1/V2 neutralizing antibodies. *J Virol* 2012,**86**:8319-8323.
461. Wu X, Changela A, O'Dell S, Schmidt SD, Pancera M, Yang Y, *et al.* Immunotypes of a quaternary site of HIV-1 vulnerability and their recognition by antibodies. *J Virol* 2011,**85**:4578-4585.
462. Pancera M, McLellan JS, Wu X, Zhu J, Changela A, Schmidt SD, *et al.* Crystal structure of PG16 and chimeric dissection with somatically related PG9: structure-function analysis of two quaternary-specific antibodies that effectively neutralize HIV-1. *J Virol* 2010,**84**:8098-8110.
463. Pejchal R, Doores KJ, Walker LM, Khayat R, Huang PS, Wang SK, *et al.* A potent and broad neutralizing antibody recognizes and penetrates the HIV glycan shield. *Science* 2011,**334**:1097-1103.
464. Walker LM, Huber M, Doores KJ, Falkowska E, Pejchal R, Julien JP, *et al.* Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature* 2011,**477**:466-470.
465. Bonsignori M, Hwang KK, Chen X, Tsao CY, Morris L, Gray E, *et al.* Analysis of a clonal lineage of HIV-1 envelope V2/V3 conformational epitope-specific broadly neutralizing antibodies and their inferred unmutated common ancestors. *J Virol* 2011,**85**:9998-10009.
466. Pietzsch J, Scheid JF, Mouquet H, Klein F, Seaman MS, Jankovic M, *et al.* Human anti-HIV-neutralizing antibodies frequently target a conserved epitope essential for viral fitness. *J Exp Med* 2010,**207**:1995-2002.
467. Balla-Jhagjhoorsingh SS, Willems B, Heyndrickx L, Heyndrickx L, Vereecken K, Janssens W, *et al.* Characterization of neutralizing profiles in HIV-1 infected patients from whom the HJ16, HGN194 and HK20 mAbs were obtained. *PLoS One* 2011,**6**:e25488.
468. Corti D, Langedijk JP, Hinz A, Seaman MS, Vanzetta F, Fernandez-Rodriguez BM, *et al.* Analysis of memory B cell responses and isolation of novel monoclonal antibodies with neutralizing breadth from HIV-1-infected individuals. *PLoS One* 2010,**5**:e8805.

469. Watkins JD, Siddappa NB, Lakhashe SK, Humbert M, Sholukh A, Hemashettar G, *et al.* An anti-HIV-1 V3 loop antibody fully protects cross-clade and elicits T-cell immunity in macaques mucosally challenged with an R5 clade C SHIV. *PLoS One* 2011,**6**:e18207.
470. Wu X, Yang ZY, Li Y, Hogerkorp CM, Schief WR, Seaman MS, *et al.* Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science* 2010,**329**:856-861.
471. West AP, Jr., Diskin R, Nussenzweig MC, Bjorkman PJ. Structural basis for germ-line gene usage of a potent class of antibodies targeting the CD4-binding site of HIV-1 gp120. *Proc Natl Acad Sci U S A* 2012,**109**:E2083-2090.
472. Zhou T, Zhu J, Wu X, Moquin S, Zhang B, Acharya P, *et al.* Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-class antibodies. *Immunity* 2013,**39**:245-258.
473. Wu X, Zhou T, Zhu J, Zhang B, Georgiev I, Wang C, *et al.* Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* 2011,**333**:1593-1602.
474. Falkowska E, Ramos A, Feng Y, Zhou T, Moquin S, Walker LM, *et al.* PGV04, an HIV-1 gp120 CD4 binding site antibody, is broad and potent in neutralization but does not induce conformational changes characteristic of CD4. *J Virol* 2012,**86**:4394-4403.
475. Jiang X, Burke V, Totrov M, Williams C, Cardozo T, Gorny MK, *et al.* Conserved structural elements in the V3 crown of HIV-1 gp120. *Nat Struct Mol Biol* 2010,**17**:955-961.
476. Gorny MK, Sampson J, Li H, Jiang X, Totrov M, Wang XH, *et al.* Human anti-V3 HIV-1 monoclonal antibodies encoded by the VH5-51/VL lambda genes define a conserved antigenic structure. *PLoS One* 2011,**6**:e27780.
477. Stanfield RL, Gorny MK, Zolla-Pazner S, Wilson IA. Crystal structures of human immunodeficiency virus type 1 (HIV-1) neutralizing antibody 2219 in complex with three different V3 peptides reveal a new binding mode for HIV-1 cross-reactivity. *J Virol* 2006,**80**:6093-6105.
478. Changela A, Wu X, Yang Y, Zhang B, Zhu J, Nardone GA, *et al.* Crystal structure of human antibody 2909 reveals conserved features of quaternary structure-specific antibodies that potently neutralize HIV-1. *J Virol* 2011,**85**:2524-2535.
479. Krachmarov C, Lai Z, Honnen WJ, Salomon A, Gorny MK, Zolla-Pazner S, *et al.* Characterization of structural features and diversity of variable-region determinants of related quaternary epitopes recognized by human and rhesus macaque monoclonal antibodies possessing unusually potent neutralizing activities. *J Virol* 2011,**85**:10730-10740.
480. Mann A, Friedrich N, Krarup A, Weber J, Stiegeler E, Dreier B, *et al.* Conformation-dependent recognition of HIV gp120 by designed ankyrin repeat proteins provides access to novel HIV entry inhibitors. *J Virol* 2013,**87**:5868-5881.
481. Stanfield RL, Gorny MK, Williams C, Zolla-Pazner S, Wilson IA. Structural rationale for the broad neutralization of HIV-1 by human monoclonal antibody 447-52D. *Structure* 2004,**12**:193-204.
482. Burke V, Williams C, Sukumaran M, Kim SS, Li H, Wang XH, *et al.* Structural basis of the cross-reactivity of genetically related human anti-HIV-1 mAbs: implications for design of V3-based immunogens. *Structure* 2009,**17**:1538-1546.
483. Killikelly A, Zhang HT, Spurrier B, Williams C, Gorny MK, Zolla-Pazner S, *et al.* Thermodynamic Signatures of the Antigen Binding Site of mAb 447-52D Targeting the Third Variable Region of HIV-1 gp120. *Biochemistry* 2013.
484. Sharpe S, Kessler N, Anglister JA, Yau WM, Tycko R. Solid-state NMR yields structural constraints on the V3 loop from HIV-1 Gp120 bound to the 447-52D antibody Fv fragment. *J Am Chem Soc* 2004,**126**:4979-4990.
485. Weliky DP, Bennett AE, Zvi A, Anglister J, Steinbach PJ, Tycko R. Solid-state NMR evidence for an antibody-dependent conformation of the V3 loop of HIV-1 gp120. *Nat Struct Biol* 1999,**6**:141-145.
486. Zvi A, Tugarinov V, Faiman GA, Horovitz A, Anglister J. A model of a gp120 V3 peptide in complex with an HIV-neutralizing antibody based on NMR and mutant cycle-derived constraints. *Eur J Biochem* 2000,**267**:767-779.
487. Zhou C, Lu L, Tan S, Jiang S, Chen YH. HIV-1 glycoprotein 41 ectodomain induces activation of the CD74 protein-mediated extracellular signal-regulated kinase/mitogen-activated protein kinase pathway to enhance viral infection. *J Biol Chem* 2011,**286**:44869-44877.

488. Zhang H, Wang L, Kao S, Whitehead IP, Hart MJ, Liu B, *et al.* Functional interaction between the cytoplasmic leucine-zipper domain of HIV-1 gp41 and p115-RhoGEF. *Curr Biol* 1999,**9**:1271-1274.
489. Blot G, Lopez-Verges S, Treand C, Kubat NJ, Delcroix-Genete D, Emiliani S, *et al.* Luman, a new partner of HIV-1 TMgp41, interferes with Tat-mediated transcription of the HIV-1 LTR. *J Mol Biol* 2006,**364**:1034-1047.
490. Kim JT, Kim EM, Lee KH, Choi JE, Jhun BH, Kim JW. Leucine zipper domain of HIV-1 gp41 interacted specifically with alpha-catenin. *Biochem Biophys Res Commun* 2002,**291**:1239-1244.
491. Song L, Sun ZY, Coleman KE, Zwick MB, Gach JS, Wang JH, *et al.* Broadly neutralizing anti-HIV-1 antibodies disrupt a hinge-related function of gp41 at the membrane interface. *Proc Natl Acad Sci U S A* 2009,**106**:9057-9062.
492. Pejchal R, Gach JS, Brunel FM, Cardoso RM, Stanfield RL, Dawson PE, *et al.* A conformational switch in human immunodeficiency virus gp41 revealed by the structures of overlapping epitopes recognized by neutralizing antibodies. *J Virol* 2009,**83**:8451-8462.
493. Srinivas SK, Srinivas RV, Anantharamaiah GM, Compans RW, Segrest JP. Cytosolic domain of the human immunodeficiency virus envelope glycoproteins binds to calmodulin and inhibits calmodulin-regulated proteins. *J Biol Chem* 1993,**268**:22895-22899.
494. Micoli KJ, Mamaeva O, Piller SC, Barker JL, Pan G, Hunter E, *et al.* Point mutations in the C-terminus of HIV-1 gp160 reduce apoptosis and calmodulin binding without affecting viral replication. *Virology* 2006,**344**:468-479.
495. Sham SW, McDonald JM, Micoli KJ, Krishna NR. Solution structure of a calmodulin-binding domain in the carboxy-terminal region of HIV type 1 gp160. *AIDS Res Hum Retroviruses* 2008,**24**:607-616.
496. Hovanessian AG, Briand JP, Said EA, Svab J, Ferris S, Dali H, *et al.* The caveolin-1 binding domain of HIV-1 glycoprotein gp41 is an efficient B cell epitope vaccine candidate against virus infection. *Immunity* 2004,**21**:617-627.
497. Speth C, Prohaszka Z, Mair M, Stockl G, Zhu X, Jobstl B, *et al.* A 60 kD heat-shock protein-like molecule interacts with the HIV transmembrane glycoprotein gp41. *Mol Immunol* 1999,**36**:619-628.
498. Gallo SA, Wang W, Rawat SS, Jung G, Waring AJ, Cole AM, *et al.* Theta-defensins prevent HIV-1 Env-mediated fusion by binding gp41 and blocking 6-helix bundle formation. *J Biol Chem* 2006,**281**:18787-18792.
499. Fausther-Bovendo H, Vieillard V, Sagan S, Bismuth G, Debre P. HIV gp41 engages gC1qR on CD4+ T cells to induce the expression of an NK ligand through the PIP3/H2O2 pathway. *PLoS Pathog* 2010,**6**:e1000975.
500. Poon DT, Coren LV, Ott DE. Efficient incorporation of HLA class II onto human immunodeficiency virus type 1 requires envelope glycoprotein packaging. *J Virol* 2000,**74**:3918-3923.
501. Alfsen A, Bomsel M. HIV-1 gp41 envelope residues 650-685 exposed on native virus act as a lectin to bind epithelial cell galactosyl ceramide. *J Biol Chem* 2002,**277**:25649-25659.
502. Postler TS, Desrosiers RC. The cytoplasmic domain of the HIV-1 glycoprotein gp41 induces NF-kappaB activation through TGF-beta-activated kinase 1. *Cell Host Microbe* 2012,**11**:181-193.
503. Ohno H, Aguilar RC, Fournier MC, Hennecke S, Cosson P, Bonifacino JS. Interaction of endocytic signals from the HIV-1 envelope glycoprotein complex with members of the adaptor medium chain family. *Virology* 1997,**238**:305-315.
504. Wyss S, Berlioz-Torrent C, Boge M, Blot G, Honing S, Benarous R, *et al.* The highly conserved C-terminal dileucine motif in the cytosolic domain of the human immunodeficiency virus type 1 envelope glycoprotein is critical for its association with the AP-1 clathrin adaptor [correction of adapter]. *J Virol* 2001,**75**:2982-2992.
505. Berlioz-Torrent C, Shacklett BL, Erdtmann L, Delamarre L, Bouchaert I, Sonigo P, *et al.* Interactions of the cytoplasmic domains of human and simian retroviral transmembrane proteins with components of the clathrin adaptor complexes modulate intracellular and cell surface expression of envelope glycoproteins. *J Virol* 1999,**73**:1350-1361.
506. Byland R, Vance PJ, Hoxie JA, Marsh M. A conserved dileucine motif mediates clathrin and AP-2-dependent endocytosis of the HIV-1 envelope protein. *Mol Biol Cell* 2007,**18**:414-425.
507. Boge M, Wyss S, Bonifacino JS, Thali M. A membrane-proximal tyrosine-based signal mediates internalization of the HIV-1 envelope glycoprotein via interaction with the AP-2 clathrin adaptor. *J Biol Chem* 1998,**273**:15773-15778.

508. Noble B, Abada P, Nunez-Iglesias J, Cannon PM. Recruitment of the adaptor protein 2 complex by the human immunodeficiency virus type 2 envelope protein is necessary for high levels of virus release. *J Virol* 2006,**80**:2924-2932.
509. Reading SA, Heap CJ, Dimmock NJ. A novel monoclonal antibody specific to the C-terminal tail of the gp41 envelope transmembrane protein of human immunodeficiency virus type 1 that preferentially neutralizes virus after it has attached to the target cell and inhibits the production of infectious progeny. *Virology* 2003,**315**:362-372.
510. Steckbeck JD, Sun C, Sturgeon TJ, Montelaro RC. Topology of the C-terminal tail of HIV-1 gp41: differential exposure of the Kennedy epitope on cell and viral membranes. *PLoS One* 2010,**5**:e15261.
511. Julien JP, Huarte N, Maeso R, Taneva SG, Cunningham A, Nieva JL, *et al.* Ablation of the complementarity-determining region H3 apex of the anti-HIV-1 broadly neutralizing antibody 2F5 abrogates neutralizing capacity without affecting core epitope binding. *J Virol* 2010,**84**:4136-4147.
512. Zwick MB, Labrijn AF, Wang M, Spenlehauer C, Saphire EO, Binley JM, *et al.* Broadly neutralizing antibodies targeted to the membrane-proximal external region of human immunodeficiency virus type 1 glycoprotein gp41. *J Virol* 2001,**75**:10892-10905.
513. Muster T, Steindl F, Purtscher M, Trkola A, Klima A, Himmler G, *et al.* A conserved neutralizing epitope on gp41 of human immunodeficiency virus type 1. *J Virol* 1993,**67**:6642-6647.
514. Montero M, Gulzar N, Klaric KA, Donald JE, Lepik C, Wu S, *et al.* Neutralizing epitopes in the membrane-proximal external region of HIV-1 gp41 are influenced by the transmembrane domain and the plasma membrane. *J Virol* 2012,**86**:2930-2941.
515. Guenaga J, Wyatt RT. Structure-guided alterations of the gp41-directed HIV-1 broadly neutralizing antibody 2F5 reveal new properties regarding its neutralizing function. *PLoS Pathog* 2012,**8**:e1002806.
516. Kim M, Sun ZY, Rand KD, Shi X, Song L, Cheng Y, *et al.* Antibody mechanics on a membrane-bound HIV segment essential for GP41-targeted viral neutralization. *Nat Struct Mol Biol* 2011,**18**:1235-1243.
517. de la Arada I, Julien JP, de la Torre BG, Huarte N, Andreu D, Pai EF, *et al.* Structural constraints imposed by the conserved fusion peptide on the HIV-1 gp41 epitope recognized by the broadly neutralizing antibody 2F5. *J Phys Chem B* 2009,**113**:13626-13637.
518. Muster T, Guinea R, Trkola A, Purtscher M, Klima A, Steindl F, *et al.* Cross-neutralizing activity against divergent human immunodeficiency virus type 1 isolates induced by the gp41 sequence ELDKWAS. *J Virol* 1994,**68**:4031-4034.
519. Ofek G, Tang M, Sambor A, Katinger H, Mascola JR, Wyatt R, *et al.* Structure and mechanistic analysis of the anti-human immunodeficiency virus type 1 antibody 2F5 in complex with its gp41 epitope. *J Virol* 2004,**78**:10724-10737.
520. Alam SM, Morelli M, Dennison SM, Liao HX, Zhang R, Xia SM, *et al.* Role of HIV membrane in neutralization by two broadly neutralizing antibodies. *Proc Natl Acad Sci U S A* 2009,**106**:20234-20239.
521. Vella C, Ferguson M, Dunn G, Meloen R, Langedijk H, Evans D, *et al.* Characterization and primary structure of a human immunodeficiency virus type 1 (HIV-1) neutralization domain as presented by a poliovirus type 1/HIV-1 chimera. *J Gen Virol* 1993,**74 (Pt 12)**:2603-2607.
522. Cardoso RM, Zwick MB, Stanfield RL, Kunert R, Binley JM, Katinger H, *et al.* Broadly neutralizing anti-HIV antibody 4E10 recognizes a helical conformation of a highly conserved fusion-associated motif in gp41. *Immunity* 2005,**22**:163-173.
523. Cardoso RM, Brunel FM, Ferguson S, Zwick M, Burton DR, Dawson PE, *et al.* Structural basis of enhanced binding of extended and helically constrained peptide epitopes of the broadly neutralizing HIV-1 antibody 4E10. *J Mol Biol* 2007,**365**:1533-1544.
524. Sun ZY, Oh KJ, Kim M, Yu J, Brusic V, Song L, *et al.* HIV-1 broadly neutralizing antibody extracts its epitope from a kinked gp41 ectodomain region on the viral membrane. *Immunity* 2008,**28**:52-63.
525. Huang J, Ofek G, Laub L, Louder MK, Doria-Rose NA, Longo NS, *et al.* Broad and potent neutralization of HIV-1 by a gp41-specific human antibody. *Nature* 2012,**491**:406-412.
526. Morris L, Chen X, Alam M, Tomaras G, Zhang R, Marshall DJ, *et al.* Isolation of a human anti-HIV gp41 membrane proximal region neutralizing antibody by antigen-specific single B cell sorting. *PLoS One* 2011,**6**:e23532.

527. Nicely NI, Dennison SM, Spicer L, Scearce RM, Kelsoe G, Ueda Y, *et al.* Crystal structure of a non-neutralizing antibody to the HIV-1 gp41 membrane-proximal external region. *Nat Struct Mol Biol* 2010,**17**:1492-1494.
528. Sabin C, Corti D, Buzon V, Seaman MS, Lutje Hulsik D, Hinz A, *et al.* Crystal structure and size-dependent neutralization properties of HK20, a human monoclonal antibody binding to the highly conserved heptad repeat 1 of gp41. *PLoS Pathog* 2010,**6**:e1001195.
529. Zhang MY, Vu BK, Choudhary A, Lu H, Humbert M, Ong H, *et al.* Cross-reactive human immunodeficiency virus type 1-neutralizing human monoclonal antibody that recognizes a novel conformational epitope on gp41 and lacks reactivity against self-antigens. *J Virol* 2008,**82**:6869-6879.
530. Gustchina E, Li M, Louis JM, Anderson DE, Lloyd J, Frisch C, *et al.* Structural basis of HIV-1 neutralization by affinity matured Fabs directed against the internal trimeric coiled-coil of gp41. *PLoS Pathog* 2010,**6**:e1001182.
531. Frey G, Chen J, Rits-Volloch S, Freeman MM, Zolla-Pazner S, Chen B. Distinct conformational states of HIV-1 gp41 are recognized by neutralizing and non-neutralizing antibodies. *Nat Struct Mol Biol* 2010,**17**:1486-1491.
532. Gustchina E, Louis JM, Lam SN, Bewley CA, Clore GM. A monoclonal Fab derived from a human nonimmune phage library reveals a new epitope on gp41 and neutralizes diverse human immunodeficiency virus type 1 strains. *J Virol* 2007,**81**:12946-12953.
533. Yang P, Ai LS, Huang SC, Li HF, Chan WE, Chang CW, *et al.* The cytoplasmic domain of human immunodeficiency virus type 1 transmembrane protein gp41 harbors lipid raft association determinants. *J Virol* 2010,**84**:59-75.
534. Bhattacharya J, Peters PJ, Clapham PR. Human immunodeficiency virus type 1 envelope glycoproteins that lack cytoplasmic domain cysteines: impact on association with membrane lipid rafts and incorporation onto budding virus particles. *J Virol* 2004,**78**:5500-5506.
535. Abacioglu YH, Fouts TR, Laman JD, Claassen E, Pincus SH, Moore JP, *et al.* Epitope mapping and topology of baculovirus-expressed HIV-1 gp160 determined with a panel of murine monoclonal antibodies. *AIDS Res Hum Retroviruses* 1994,**10**:371-381.
536. Cleveland SM, Buratti E, Jones TD, North P, Baralle F, McLain L, *et al.* Immunogenic and antigenic dominance of a nonneutralizing epitope over a highly conserved neutralizing epitope in the gp41 envelope glycoprotein of human immunodeficiency virus type 1: its deletion leads to a strong neutralizing response. *Virology* 2000,**266**:66-78.
537. Emerson V, Holtkotte D, Pfeiffer T, Wang IH, Schnolzer M, Kempf T, *et al.* Identification of the cellular prohibitin 1/prohibitin 2 heterodimer as an interaction partner of the C-terminal cytoplasmic domain of the HIV-1 glycoprotein. *J Virol* 2010,**84**:1355-1365.
538. Bultmann A, Muranyi W, Seed B, Haas J. Identification of two sequences in the cytoplasmic tail of the human immunodeficiency virus type 1 envelope glycoprotein that inhibit cell surface expression. *J Virol* 2001,**75**:5263-5276.
539. Henderson LA, Qureshi MN. A peptide inhibitor of human immunodeficiency virus infection binds to novel human cell surface polypeptides. *J Biol Chem* 1993,**268**:15291-15297.
540. Xiao Y, Wu W, Dierich MP, Chen Y. HIV-1 gp41 by N-domain binds the potential receptor protein P45. *Int Arch Allergy Immunol* 2000,**121**:253-257.
541. Trible RP, Emert-Sedlak L, Smithgall TE. HIV-1 Nef selectively activates Src family kinases Hck, Lyn, and c-Src through direct SH3 domain interaction. *J Biol Chem* 2006,**281**:27029-27038.
542. Lee CH, Saksela K, Mirza UA, Chait BT, Kuriyan J. Crystal structure of the conserved core of HIV-1 Nef complexed with a Src family SH3 domain. *Cell* 1996,**85**:931-942.
543. Pene-Dumitrescu T, Shu ST, Wales TE, Alvarado JJ, Shi H, Narute P, *et al.* HIV-1 Nef interaction influences the ATP-binding site of the Src-family kinase, Hck. *BMC Chem Biol* 2012,**12**:1.
544. Arold S, Franken P, Strub MP, Hoh F, Benichou S, Benarous R, *et al.* The crystal structure of HIV-1 Nef protein bound to the Fyn kinase SH3 domain suggests a role for this complex in altered T cell receptor signaling. *Structure* 1997,**5**:1361-1372.
545. Briggs SD, Lerner EC, Smithgall TE. Affinity of Src family kinase SH3 domains for HIV Nef in vitro does not predict kinase activation by Nef in vivo. *Biochemistry* 2000,**39**:489-495.
546. Saksela K, Cheng G, Baltimore D. Proline-rich (PxxP) motifs in HIV-1 Nef bind to SH3 domains of a subset of Src kinases and are required for the enhanced growth of Nef+ viruses but not for down-regulation of CD4. *EMBO J* 1995,**14**:484-491.

547. Hung CH, Thomas L, Ruby CE, Atkins KM, Morris NP, Knight ZA, *et al.* HIV-1 Nef assembles a Src family kinase-ZAP-70/Syk-PI3K cascade to downregulate cell-surface MHC-I. *Cell Host Microbe* 2007,**1**:121-133.
548. Lee CH, Leung B, Lemmon MA, Zheng J, Cowburn D, Kuriyan J, *et al.* A single amino acid in the SH3 domain of Hck determines its high affinity and specificity in binding to HIV-1 Nef protein. *EMBO J* 1995,**14**:5006-5015.
549. Kuo LS, Baugh LL, Denial SJ, Watkins RL, Liu M, Garcia JV, *et al.* Overlapping effector interfaces define the multiple functions of the HIV-1 Nef polyproline helix. *Retrovirology* 2012,**9**:47.
550. Jung J, Byeon IJ, Ahn J, Gronenborn AM. Structure, dynamics, and Hck interaction of full-length HIV-1 Nef. *Proteins* 2011,**79**:1609-1622.
551. Greenway A, Azad A, Mills J, McPhee D. Human immunodeficiency virus type 1 Nef binds directly to Lck and mitogen-activated protein kinase, inhibiting kinase activity. *J Virol* 1996,**70**:6701-6708.
552. Hill BT, Skowronski J. Human N-myristoyltransferases form stable complexes with lentiviral nef and other viral and cellular substrate proteins. *J Virol* 2005,**79**:1133-1141.
553. Morgan CR, Miglionico BV, Engen JR. Effects of HIV-1 Nef on human N-myristoyltransferase 1. *Biochemistry* 2011,**50**:3394-3403.
554. Fackler OT, Lu X, Frost JA, Geyer M, Jiang B, Luo W, *et al.* p21-activated kinase 1 plays a critical role in cellular activation by Nef. *Mol Cell Biol* 2000,**20**:2619-2627.
555. Grzesiek S, Stahl SJ, Wingfield PT, Bax A. The CD4 determinant for downregulation by HIV-1 Nef directly binds to Nef. Mapping of the Nef binding surface by NMR. *Biochemistry* 1996,**35**:10256-10261.
556. Greenberg M, DeTulleo L, Rapoport I, Skowronski J, Kirchhausen T. A dileucine motif in HIV-1 Nef is essential for sorting into clathrin-coated pits and for downregulation of CD4. *Curr Biol* 1998,**8**:1239-1242.
557. Swigut T, Shohdy N, Skowronski J. Mechanism for down-regulation of CD28 by Nef. *EMBO J* 2001,**20**:1593-1604.
558. Geyer M, Peterlin BM. Domain assembly, surface accessibility and sequence conservation in full length HIV-1 Nef. *FEBS Lett* 2001,**496**:91-95.
559. Geyer M, Fackler OT, Peterlin BM. Subunit H of the V-ATPase involved in endocytosis shows homology to beta-adaptins. *Mol Biol Cell* 2002,**13**:2045-2056.
560. Geyer M, Yu H, Mandic R, Linnemann T, Zheng YH, Fackler OT, *et al.* Subunit H of the V-ATPase binds to the medium chain of adaptor protein complex 2 and connects Nef to the endocytic machinery. *J Biol Chem* 2002,**277**:28521-28529.
561. Mujawar Z, Tamehiro N, Grant A, Sviridov D, Bukrinsky M, Fitzgerald ML. Mutation of the ATP cassette binding transporter A1 (ABCA1) C-terminus disrupts HIV-1 Nef binding but does not block the Nef enhancement of ABCA1 protein degradation. *Biochemistry* 2010,**49**:8338-8349.
562. Hodge DR, Dunn KJ, Pei GK, Chakrabarty MK, Heidecker G, Lautenberger JA, *et al.* Binding of c-Raf1 kinase to a conserved acidic sequence within the carboxyl-terminal region of the HIV-1 Nef protein. *J Biol Chem* 1998,**273**:15727-15733.
563. Fackler OT, Luo W, Geyer M, Alberts AS, Peterlin BM. Activation of Vav by Nef induces cytoskeletal rearrangements and downstream effector functions. *Mol Cell* 1999,**3**:729-739.
564. Rauch S, Pulkkinen K, Saksela K, Fackler OT. Human immunodeficiency virus type 1 Nef recruits the guanine exchange factor Vav1 via an unexpected interface into plasma membrane microdomains for association with p21-activated kinase 2 activity. *J Virol* 2008,**82**:2918-2929.
565. Lu TC, He JC, Wang ZH, Feng X, Fukumi-Tominaga T, Chen N, *et al.* HIV-1 Nef disrupts the podocyte actin cytoskeleton by interacting with diaphanous interacting protein. *J Biol Chem* 2008,**283**:8173-8182.
566. Gallina A, Rossi F, Milanesi G. Rack1 binds HIV-1 Nef and can act as a Nef-protein kinase C adaptor. *Virology* 2001,**283**:7-18.
567. Linnemann T, Zheng YH, Mandic R, Peterlin BM. Interaction between Nef and phosphatidylinositol-3-kinase leads to activation of p21-activated kinase and increased production of HIV. *Virology* 2002,**294**:246-255.
568. Kumar B, Tripathi C, Kanchan RK, Tripathi JK, Ghosh JK, Ramachandran R, *et al.* Dynamics of physical interaction between HIV-1 Nef and ASK1: identifying the interacting motif(s). *PLoS One* 2013,**8**:e67586.

569. Geleziunas R, Xu W, Takeda K, Ichijo H, Greene WC. HIV-1 Nef inhibits ASK1-dependent death signalling providing a potential mechanism for protecting the infected host cell. *Nature* 2001,**410**:834-838.
570. Baugh LL, Garcia JV, Foster JL. Functional characterization of the human immunodeficiency virus type 1 Nef acidic domain. *J Virol* 2008,**82**:9657-9667.
571. Piguet V, Wan L, Borel C, Mangasarian A, Demaurex N, Thomas G, *et al.* HIV-1 Nef protein binds to the cellular protein PACS-1 to downregulate class I major histocompatibility complexes. *Nat Cell Biol* 2000,**2**:163-167.
572. Blagoveshchenskaya AD, Thomas L, Feliciangeli SF, Hung CH, Thomas G. HIV-1 Nef downregulates MHC-I by a PACS-1- and PI3K-regulated ARF6 endocytic pathway. *Cell* 2002,**111**:853-866.
573. Dikeakos JD, Thomas L, Kwon G, Elferich J, Shinde U, Thomas G. An interdomain binding site on HIV-1 Nef interacts with PACS-1 and PACS-2 on endosomes to down-regulate MHC-I. *Mol Biol Cell* 2012,**23**:2184-2197.
574. Atkins KM, Thomas L, Youker RT, Harriff MJ, Pissani F, You H, *et al.* HIV-1 Nef binds PACS-2 to assemble a multikinase cascade that triggers major histocompatibility complex class I (MHC-I) down-regulation: analysis using short interfering RNA and knock-out mice. *J Biol Chem* 2008,**283**:11772-11784.
575. Jia X, Singh R, Homann S, Yang H, Guatelli J, Xiong Y. Structural basis of evasion of cellular adaptive immunity by HIV-1 Nef. *Nat Struct Mol Biol* 2012,**19**:701-706.
576. Iijima S, Lee YJ, Ode H, Arold ST, Kimura N, Yokoyama M, *et al.* A noncanonical mu-1A-binding motif in the N terminus of HIV-1 Nef determines its ability to downregulate major histocompatibility complex class I in T lymphocytes. *J Virol* 2012,**86**:3944-3951.
577. Bresnahan PA, Yonemoto W, Ferrell S, Williams-Herman D, Geleziunas R, Greene WC. A dileucine motif in HIV-1 Nef acts as an internalization signal for CD4 downregulation and binds the AP-1 clathrin adaptor. *Curr Biol* 1998,**8**:1235-1238.
578. Erdtmann L, Janvier K, Raposo G, Craig HM, Benaroch P, Berlioz-Torrent C, *et al.* Two independent regions of HIV-1 Nef are required for connection with the endocytic pathway through binding to the mu 1 chain of AP1 complex. *Traffic* 2000,**1**:871-883.
579. Singh RK, Lau D, Noviello CM, Ghosh P, Guatelli JC. An MHC-I cytoplasmic domain/HIV-1 Nef fusion protein binds directly to the mu subunit of the AP-1 endosomal coat complex. *PLoS One* 2009,**4**:e8364.
580. Noviello CM, Benichou S, Guatelli JC. Cooperative binding of the class I major histocompatibility complex cytoplasmic domain and human immunodeficiency virus type 1 Nef to the endosomal AP-1 complex via its mu subunit. *J Virol* 2008,**82**:1249-1258.
581. Coleman SH, Van Damme N, Day JR, Noviello CM, Hitchin D, Madrid R, *et al.* Leucine-specific, functional interactions between human immunodeficiency virus type 1 Nef and adaptor protein complexes. *J Virol* 2005,**79**:2066-2078.
582. Roeth JF, Williams M, Kasper MR, Filzen TM, Collins KL. HIV-1 Nef disrupts MHC-I trafficking by recruiting AP-1 to the MHC-I cytoplasmic tail. *J Cell Biol* 2004,**167**:903-913.
583. Janvier K, Kato Y, Boehm M, Rose JR, Martina JA, Kim BY, *et al.* Recognition of dileucine-based sorting signals from HIV-1 Nef and LIMP-II by the AP-1 gamma-sigma1 and AP-3 delta-sigma3 hemicomplexes. *J Cell Biol* 2003,**163**:1281-1290.
584. Janvier K, Craig H, Hitchin D, Madrid R, Sol-Foulon N, Renault L, *et al.* HIV-1 Nef stabilizes the association of adaptor protein complexes with membranes. *J Biol Chem* 2003,**278**:8725-8732.
585. Lindwasser OW, Smith WJ, Chaudhuri R, Yang P, Hurley JH, Bonifacino JS. A diacidic motif in human immunodeficiency virus type 1 Nef is a novel determinant of binding to AP-2. *J Virol* 2008,**82**:1166-1174.
586. Chaudhuri R, Mattera R, Lindwasser OW, Robinson MS, Bonifacino JS. A basic patch on alpha-adaptin is required for binding of human immunodeficiency virus type 1 Nef and cooperative assembly of a CD4-Nef-AP-2 complex. *J Virol* 2009,**83**:2518-2530.
587. Chaudhuri R, Lindwasser OW, Smith WJ, Hurley JH, Bonifacino JS. Downregulation of CD4 by human immunodeficiency virus type 1 Nef is dependent on clathrin and involves direct interaction of Nef with the AP2 clathrin adaptor. *J Virol* 2007,**81**:3877-3890.
588. Mitchell RS, Chaudhuri R, Lindwasser OW, Tanaka KA, Lau D, Murillo R, *et al.* Competition model for upregulation of the major histocompatibility complex class II-associated invariant chain by human immunodeficiency virus type 1 Nef. *J Virol* 2008,**82**:7758-7767.

589. Craig HM, Reddy TR, Riggs NL, Dao PP, Guatelli JC. Interactions of HIV-1 nef with the mu subunits of adaptor protein complexes 1, 2, and 3: role of the dileucine-based sorting motif. *Virology* 2000,**271**:9-17.
590. Schaefer MR, Wonderlich ER, Roeth JF, Leonard JA, Collins KL. HIV-1 Nef targets MHC-I and CD4 for degradation via a final common beta-COP-dependent pathway in T cells. *PLoS Pathog* 2008,**4**:e1000131.
591. Piguet V, Gu F, Foti M, Demaurex N, Gruenberg J, Carpentier JL, *et al.* Nef-induced CD4 degradation: a diacidic-based motif in Nef functions as a lysosomal targeting signal through the binding of beta-COP in endosomes. *Cell* 1999,**97**:63-73.
592. Benichou S, Bomsel M, Bodeus M, Durand H, Doute M, Letourneur F, *et al.* Physical interaction of the HIV-1 Nef protein with beta-COP, a component of non-clathrin-coated vesicles essential for membrane traffic. *J Biol Chem* 1994,**269**:30073-30076.
593. Agopian K, Wei BL, Garcia JV, Gabuzda D. A hydrophobic binding surface on the human immunodeficiency virus type 1 Nef core is critical for association with p21-activated kinase 2. *J Virol* 2006,**80**:3050-3061.
594. Stolp B, Abraham L, Rudolph JM, Fackler OT. Lentiviral Nef proteins utilize PAK2-mediated deregulation of cofilin as a general strategy to interfere with actin remodeling. *J Virol* 2010,**84**:3935-3948.
595. Olivieri KC, Mukerji J, Gabuzda D. Nef-mediated enhancement of cellular activation and human immunodeficiency virus type 1 replication in primary T cells is dependent on association with p21-activated kinase 2. *Retrovirology* 2011,**8**:64.
596. Costa LJ, Chen N, Lopes A, Aguiar RS, Tanuri A, Plemenitas A, *et al.* Interactions between Nef and AIP1 proliferate multivesicular bodies and facilitate egress of HIV-1. *Retrovirology* 2006,**3**:33.
597. Kyei GB, Dinkins C, Davis AS, Roberts E, Singh SB, Dong C, *et al.* Autophagy pathway intersects with HIV-1 biosynthesis and regulates viral yields in macrophages. *J Cell Biol* 2009,**186**:255-268.
598. Shelton MN, Huang MB, Ali SA, Powell MD, Bond VC. Secretion modification region-derived peptide disrupts HIV-1 Nef's interaction with mortalin and blocks virus and Nef exosome release. *J Virol* 2012,**86**:406-419.
599. Kumar M, Rawat P, Khan SZ, Dhamija N, Chaudhary P, Ravi DS, *et al.* Reciprocal regulation of human immunodeficiency virus-1 gene expression and replication by heat shock proteins 40 and 70. *J Mol Biol* 2011,**410**:944-958.
600. Matsubara M, Jing T, Kawamura K, Shimojo N, Titani K, Hashimoto K, *et al.* Myristoyl moiety of HIV Nef is involved in regulation of the interaction with calmodulin in vivo. *Protein Sci* 2005,**14**:494-503.
601. Chandrasekaran P, Buckley M, Moore V, Wang LQ, Kehrl JH, Venkatesan S. HIV-1 Nef impairs heterotrimeric G-protein signaling by targeting Galpha(i2) for degradation through ubiquitination. *J Biol Chem* 2012,**287**:41481-41498.
602. Xu XN, Laffert B, Screaton GR, Kraft M, Wolf D, Kolanus W, *et al.* Induction of Fas ligand expression by HIV involves the interaction of Nef with the T cell receptor zeta chain. *J Exp Med* 1999,**189**:1489-1496.
603. Cohen GB, Rangan VS, Chen BK, Smith S, Baltimore D. The human thioesterase II protein binds to a site on HIV-1 Nef critical for CD4 down-regulation. *J Biol Chem* 2000,**275**:23097-23105.
604. Liu LX, Heveker N, Fackler OT, Arold S, Le Gall S, Janvier K, *et al.* Mutation of a conserved residue (D123) required for oligomerization of human immunodeficiency virus type 1 Nef protein abolishes interaction with human thioesterase and results in impairment of Nef biological functions. *J Virol* 2000,**74**:5310-5319.
605. Janardhan A, Swigut T, Hill B, Myers MP, Skowronski J. HIV-1 Nef binds the DOCK2-ELMO1 complex to activate rac and inhibit lymphocyte chemotaxis. *PLoS Biol* 2004,**2**:E6.
606. Greenway AL, McPhee DA, Allen K, Johnstone R, Holloway G, Mills J, *et al.* Human immunodeficiency virus type 1 Nef binds to tumor suppressor p53 and protects cells against p53-mediated apoptosis. *J Virol* 2002,**76**:2692-2702.
607. Baur AS, Sass G, Laffert B, Willbold D, Cheng-Mayer C, Peterlin BM. The N-terminus of Nef from HIV-1/SIV associates with a protein complex containing Lck and a serine kinase. *Immunity* 1997,**6**:283-291.
608. Weiser K, Barton M, Gershoony D, Dasgupta R, Cardozo T. HIV's Nef Interacts with beta-Catenin of the Wnt Signaling Pathway in HEK293 Cells. *PLoS One* 2013,**8**:e77865.

609. Pizzato M, Helander A, Popova E, Calistri A, Zamborlini A, Palu G, *et al.* Dynamin 2 is required for the enhancement of HIV-1 infectivity by Nef. *Proc Natl Acad Sci U S A* 2007,**104**:6812-6817.
610. Wolf D, Giese SI, Witte V, Krautkramer E, Trapp S, Sass G, *et al.* Novel (n)PKC kinases phosphorylate Nef for increased HIV transcription, replication and perinuclear targeting. *Virology* 2008,**370**:45-54.
611. Smith BL, Krushelnycky BW, Mochly-Rosen D, Berg P. The HIV nef protein associates with protein kinase C theta. *J Biol Chem* 1996,**271**:16753-16757.
612. Witte V, Laffert B, Rosorius O, Lischka P, Blume K, Galler G, *et al.* HIV-1 Nef mimics an integrin receptor signal that recruits the polycomb group protein Eed to the plasma membrane. *Mol Cell* 2004,**13**:179-190.
613. Aqil M, Naqvi AR, Bano AS, Jameel S. The HIV-1 Nef protein binds argonaute-2 and functions as a viral suppressor of RNA interference. *PLoS One* 2013,**8**:e74472.
614. Racape J, Connan F, Hoebeke J, Choppin J, Guillet JG. Influence of dominant HIV-1 epitopes on HLA-A3/peptide complex formation. *Proc Natl Acad Sci U S A* 2006,**103**:18208-18213.
615. Wonderlich ER, Williams M, Collins KL. The tyrosine binding pocket in the adaptor protein 1 (AP-1) mu1 subunit is necessary for Nef to recruit AP-1 to the major histocompatibility complex class I cytoplasmic tail. *J Biol Chem* 2008,**283**:3011-3022.
616. Williams M, Roeth JF, Kasper MR, Filzen TM, Collins KL. Human immunodeficiency virus type 1 Nef domains required for disruption of major histocompatibility complex class I trafficking are also necessary for coprecipitation of Nef with HLA-A2. *J Virol* 2005,**79**:632-636.
617. Toussaint H, Gobert FX, Schindler M, Banning C, Kozik P, Jouve M, *et al.* Human immunodeficiency virus type 1 nef expression prevents AP-2-mediated internalization of the major histocompatibility complex class II-associated invariant chain. *J Virol* 2008,**82**:8373-8382.
618. Stumptner-Cuvelette P, Morchoisne S, Dugast M, Le Gall S, Raposo G, Schwartz O, *et al.* HIV-1 Nef impairs MHC class II antigen presentation and surface expression. *Proc Natl Acad Sci U S A* 2001,**98**:12144-12149.
619. Fukushi M, Dixon J, Kimura T, Tsurutani N, Dixon MJ, Yamamoto N. Identification and cloning of a novel cellular protein Naf1, Nef-associated factor 1, that increases cell surface CD4 expression. *FEBS Lett* 1999,**442**:83-88.
620. Mangino G, Percario ZA, Fiorucci G, Vaccari G, Acconcia F, Chiarabelli C, *et al.* HIV-1 Nef induces proinflammatory state in macrophages through its acidic cluster domain: involvement of TNF alpha receptor associated factor 2. *PLoS One* 2011,**6**:e22982.
621. Fackler OT, Kienzle N, Kremmer E, Boese A, Schramm B, Klimkait T, *et al.* Association of human immunodeficiency virus Nef protein with actin is myristoylation dependent and influences its subcellular localization. *Eur J Biochem* 1997,**247**:843-851.
622. Bentham M, Mazaleyrat S, Harris M. Role of myristoylation and N-terminal basic residues in membrane association of the human immunodeficiency virus type 1 Nef protein. *J Gen Virol* 2006,**87**:563-571.
623. Gerlach H, Laumann V, Martens S, Becker CF, Goody RS, Geyer M. HIV-1 Nef membrane association depends on charge, curvature, composition and sequence. *Nat Chem Biol* 2010,**6**:46-53.
624. Akgun B, Satija S, Nanda H, Pirrone GF, Shi X, Engen JR, *et al.* Conformational Transition of Membrane-Associated Terminally Acylated HIV-1 Nef. *Structure* 2013,**21**:1822-1833.

Name: Guangdi Li

Email: liguangdi.research@gmail.com

Birth: Changsha, China

Skill: data visualization, programming.

Education

2009-2014 Ph.D. in Biomedical Sciences, Rega Institute for Medical Research, Department of Microbiology and Immunology, KU Leuven, Belgium

2010-2011 Study in Master of Molecular and Cellular Biophysics, Department of Chemistry, KU Leuven (60 credits).

2007-2008 Research internship, Department of Computer Science, Universidad Politecnica De Madrid, Spain

2006-2009 M.Sc. in Computer Science, Department of Computer Science, Shandong University, China

2002-2006 B.Sc. in Applied Mathematics, Department of Mathematics and Economics, Hunan University, China

Honors and Awards

2007-2008 Scholarship awarded by Universidad Politecnica De Madrid, Spain

2008 The first prize of China Postgraduate Mathematical Contest in Modeling

2004 The first prize of Mathematical Contest in Modeling, Hunan University

2002-2005 Scholarships awarded by Hunan University

Publication list

A. International Journals or full research articles in international conferences

1. Yanwei Wang, Zhifei Ji, **Guangdi Li**, *Energy saving design for central air-conditioning system*, Mathematics in Practice and Theory, 16, 2009.
2. Concha Bielza, **Guangdi Li,** Pedro Larrañaga, *Multi-Dimensional Classification with Bayesian Networks*, International Journal of Approximate Reasoning , 52:6, 705–727, 2011 (IF=1.7, Citation = 52)

3. **Guangdi Li,** Jens Verheyen, Soo-Yon Rhee, Arnout Voet, Anne-Mieke Vandamme, Kristof Theys, *Functional conservation of HIV-1 Gag: implications for rational drug design*. Retrovirology 10(1):126. 2013 (IF=5.66)
4. **Guangdi Li**, Anne-Mieke Vandamme, Jan Ramon, *Learning Ancestral Polytrees*. The 1st workshop of Learning Tractable Probabilistic Model co-located with the 31st International Conference on Machine Learning (ICML 2014), Beijing, China, 2014
5. **Guangdi Li,** *The complexity of probabilistic inference in multi-dimensional Bayesian classifiers*, ACSIT, Zurich Switzerland, 2014
6. **Guangdi Li**, Jens Verheyen, Kristof Theys, Supinya Piampongsant, Kristel Van Laethem, Anne-Mieke Vandamme. *HIV-1 Gag C-terminal amino acid substitutions emerging under selective pressure of protease inhibitors in patient populations infected with different HIV-1 subtypes*. Retrovirology (accepted), 2014 (IF=4.77)
7. Sarah Megens, Dolores Vaira, Greet De Baets, Nathalie Dekeersmaeker, Yoeri Schrooten, **Guangdi Li**, Joost Schymkowitz, Frederic Rousseau, Anne-Mieke Vandamme, Michel Moutschen, Kristel Van Laethem. *Horizontal gene transfer from human host to HIV-1 reverse transcriptase confers drug resistance and partly compensates for replication deficits*, Virology, 456–457:310–318, 2014 (IF=3.27)
8. Andrea-Clemencia Pineda, Nuno Rodrigues Faria, Francisco-Javier Diaz, Patricia Olaya, Casper Møller Frederiksen, **Guangdi Li,** Arley Gomez-Lopez, Philippe Lemey, Anne-Mieke Vandamme. *The Colombian epidemic is dominated by HIV-1 subtype B: a molecular epidemiology and phylodynamic study*. PLOS One, PLoS ONE 01; 9(7):e101738, 2014 (IF=3.73)

B. Research articles under review or to be submitted

1. **Guangdi Li,** Supinya Piampongsant, Nuno Rodrigues Faria, Arnout Voet, Andrea-Clemencia Pineda-Peña, Ricardo Khouri, Philippe Lemey, Anne-Mieke Vandamme, Kristof Theys. *An integrated map of HIV genome-wide diversity from a population perspective*. Under review (Retrovirology, IF=4.77), 2014.
2. **Guangdi Li**, Kristof Theys,Jens Verheyen, Andrea-Clemencia Pineda-Peña, Ricardo Khouri, Supinya Piampongsant, Mónica Eusébio, Jan Ramon, Anne-Mieke Vandamme. *A new ensemble coevolution system for detecting HIV-1 protein coevolution.* Under review (Biology Direct, IF=4.04), 2014
3. **Guangdi Li**, Guangdi Li, Jens Verheyen, Kristof Theys, Andrea-Clemencia Pineda-Peña, Ricardo Khouri, Kristel Van Laethem, Jan Ramon and Anne-Mieke Vandamme, *HIV-1 Gag and protease coevolution networks.* To be submitted, 2014.

4. **Guangdi Li**, Kristof Theys, Jan Ramon, Anne-Mieke Vandamme, *Modeling drug genetic barrier in HIV-1 patient populations by statistical free energy models*. To be submitted, 2014.
5. **Guangdi Li**, Erik De Clercq, *HIV genome-wide protein interaction*. To be submitted, 2014.
6. **Guangdi Li**, *Learning multi-polytree graphical models*. To be submitted, 2014.
7. Eline Boons, **Guangdi Li**, Els Vanstreels, Thomas Vercruysse, Christophe Pannecouque, Anne-Mieke Vandamme, Dirk Daelemans. *A stably expressed llama single-domain intrabody targeting Rev displays broad-spectrum anti-HIV activity*. Minor revision (Antiviral Research, IF=3.43), 2014.
8. Andrea-Clemencia Pineda- Peña, Nuno Rodrigues Faria, Francisco-Javier Diaz, Patricia Olaya, **Guangdi Li,** Daniela Vanegas-Otalvaro, Casper Møller Frederiksen, Arley Gomez-Lopez, Dimitrios Paraskevis, Philippe Lemey and Anne-Mieke Vandamme, *The Colombian HIV-1 epidemic is dominated by subtype B and linked with Spain*. Under review (AIDS, IF=6.55), 2014.

C. Oral presentation or published abstracts in international conferences

1. **Guangdi Li**, Concha Bielza, Pedro Larrañaga, *Learning causal polytree structure for HIV mutation patterns*, 15th International Bio-Informatics Workshop on Virus Evolution and Molecular Epidemiology, Rotterdam, The Netherlands, September 2009.
2. **Guangdi Li**, Gertjan Beheydt, Concha Bielza, Pedro Larrañaga, Ricardo Camacho,Zehava Grossman, Carlo Torti, Maurizio Zazzi, Mattia Prosperi, Rolf Kaiser, Kristel Van Laethem, Marc De Maeyer, Anne-Mieke Vandamme. *Learning ancestral polytrees for HIV-1 mutation pathways against protease inhibitor Nelfinavir*. The 9th ECCB, Gent, Belgium, September, 2010.
3. **Guangdi Li**, Gertjan beheydt, Ricardo Camacho, Anne-Mieke Vandamme, Kristof Theys. *Analysis of HIV-1 resistance pathways using multi-polytree: implications of multiple HIV-1 mutation pathways*, The 10th MEEGID, Amsterdam, The Netherlands, November, 2010 ***(Oral presentation).***
4. **Guangdi Li**, Gertjan Beheydt, Concha Bielza, Pedro Larrañaga, Ricardo Camacho,Zehava Grossman, Carlo Torti, Maurizio Zazzi, Mattia Prosperi, Rolf Kaiser, Kristel Van Laethem, Marc De Maeyer, Anne-Mieke Vandamme, *Learning ancestral polytrees for HIV-1 mutation pathways against nelfinavir*, IAP meeting, Ghent, Beligum, December 2010.
5. **Guangdi Li,** Jens Verheyen, Soo-Yon Rhee, Arnout Voet, Mónica Eusébio, Anne-Mieke Vandamme, Kristof Theys. *Functional conservation of HIV-1 gag:*

implications for rational drug design. The 14th European AIDS Conference, Brussels, Belgium, October, 2013.

6. **Guangdi Li,** Jens Verheyen, Jan Ramon, Mónica Eusébio, Kristof Theys, Anne-Mieke Vandamme, *The HIV-1 Gag and protease coevolution networks*. The 14th European AIDS Conference, Brussels, Belgium, October, 2013.
7. **Guangdi Li,** Supinya Piampongsant, Philippe Lemey, Anne-Mieke Vandamme, Kristof Theys. *Quantification of HIV genome-wide diversity*, Infectious Disease Genomics & Global Health conference, Cambridge, UK, October, 2013 ***(Oral presentation).***
8. **Guangdi Li,** Anne-Mieke Vandamme, Jan Ramon, *Learning Ancestral Polytrees*. The workshop of Learning Tractable Probabilistic Model, Beijing, China, June 25-26, 2014 ***(Oral presentation).***
9. Andrea-Clemencia Pineda-Peña, Nuno Rodrigues Faria, Francisco-Javier Diaz, Patricia Olaya, Casper Møller Frederiksen, **Guangdi Li,** Arley Gomez-Lopez, Philippe Lemey, Anne-Mieke Vandamme, *The colombian epidemic is dominated by HIV-1 subtype B: a molecular epidemiology and phylodynamic study*. The 20th International HIV Dynamics and Evolution conference, Utrecht, The Netherland, 2013.
10. Mónica Eusébio, Raf Winand, Andrea-Clemencia Pineda-Peña, **Guangdi Li**, Kristel Van Laethem, Ricardo Camacho, Anne-Mieke Vandamme, Ana Abecasis. *Significantly different TDR levels and mutation patterns in HIV-1 subtype G compared to subtype B in Portugal*. The 14th European AIDS Conference, Brussels, Belgium, October, 2013.
11. Mónica Eusébio, Raf Winand, Andrea-Clemencia Pineda-Peña, **Guangdi Li**, Kristel Van Laethem, Ricardo Camacho, Anne-Mieke Vandamme, Ana Abecasis. *Transmission clusters of drug resistance in subtype B in Portugal*. The 12th European Meeting on HIV & Hepatitis - Treatment Strategies & Antiviral Drug Resistance, Barcelona, Spain on 26 - 28 March 2014.
12. Andrea-Clemencia Pineda-Peña, Yoeri Schrooten, Lore Vinken, Fossie Ferreira, **Guangdi Li**, Nídia Sequeira Trovão, Ricardo Khouri , Inge Derdelinckx, Paul De Munter, Claudia Kücherer, Leondios G. Kostrikis, Claus Nielsen, Kirsi Liitsola, Claudia Balotta, Annemarie Wensing ,Maja Stanojevic, Roger Paredes, Jan Albert, Charles Boucher, Arley Gomez-Lopez, Eric Van Wijngaerden, Marc Van Ranst, Jurgen Vercauteren, Kristel Van Laethem ,Anne-Mieke Vandamme. *Predictors of HIV-1 transmitted drug resistance and transmission networks in the Leuven cohort.*

The 12th European Meeting on HIV & Hepatitis - Treatment Strategies & Antiviral Drug Resistance, Barcelona, Spain on 26 - 28 March 2014.